REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES
	21 April 1997	Final, 11/96 – 05/97

4. TITLE AND SUBTITLE
Cognition Models for Visual Target Discrimination

5. 5. FUNDING NUMBERS
C: DAAE07-97-C-X024

6. AUTHORS
Gary Witus

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Turing Associates, Inc.
1392 Honey Run Drive
Ann Arbor, MI 48103

8. PERFORMING ORGANIZATION REPORT NUMBER
Turing 97-1

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Tank-automotive and Armaments Command
Warren, MI 48397

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This paper reports the results of a Phase I SBIR project to identify image processing algorithms for use in modeling human visual discrimination of military ground vehicles. The approach assumes that the front-end of human visual perception is modeled by two-stage luminance and color-opponent processing, followed by multi-resolution spatial filtering. The report examines candidate algorithms for texture and intensity segmentation, shape perception, and subsequent target categorization in the target discrimination process.

14. SUBJECT TERMS
Vision modeling Target Acquisition Human Perception
Computational Vision Models

15. NUMBER OF PAGES
57

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	

NSN 7540-01-280-5500 Computer Generated STANDARD FORM 298 (Rev 2-89)
Prescribed by ANSI Std 239-18
298-102

ACKNOWLEDGEMENTS

This report was conducted with the guidance and participation of TARDEC engineers and scientists including Dr. Rob Karlsen, Dr. Grant Gerhart, Dr. Tom Meitzler, and Mr. Richard Goetz. TARDEC computer facilities and software were made available for use on the project. In particular, we made use of elements of the TARDEC National Automotive Center Visual Perception Model (NAC-VPM). The NAC-VPM was developed by OptiMetrics, Inc., of Ann Arbor MI, under contract DAAE07-94-C-R111 to the US Army TARDEC.

CONTENTS

SECTION I

I.1 Introduction

This report documents progress and results of Phase I Small Business Innovative Research project, "Cognition Models for Visual Target Discrimination," contract number DAAE07-97-C-X024, under TARDEC basic research SBIR topic A96-097, "Vision Research and Human Perception for Target Detection."

I.1.1 Background

TARDEC's mission encompasses the development and integration of ground vehicle technologies, including computer-aided design and analysis tools to facilitate cost-effective vehicle development. The TARDEC visual perception modeling initiative is directly related to camouflage, concealment, and deception technologies for combat vehicle survivability. TARDEC's charter also includes technology transfer between the military and commercial automotive sectors, and development of dual-need technologies. TARDEC's visual perception modeling initiative includes cooperative research and development agreements with major commercial automotive manufacturers to adapt military vision models to produce metrics and computer-aided engineering tools for use in automotive safety and visibility design and evaluation. As digital systems, computerized "intelligent" assistants, and driver's enhanced-vision systems become more prevalent on the battlefield and in commercial automobiles, there is an increasing need for tools and data to objectively evaluate the human factors associated with console displays and symbols. These human interface issues also fall squarely within TARDEC's responsibility, and offer potential for bi-directional technology transfer between the military and commercial automotive sectors.

TARDEC has sponsored the development of visual perception models to fill a variety of needs: (1) military applications (e.g., the evaluation of camouflage, concealment, and deception technologies); (2) commercial automotive safety applications (e.g., the evaluation of highway and automotive signals, warnings, and indicators); and (3) dual-need applications (e.g., enhanced vision systems for driving and/or target acquisition, and the design and evaluation of console displays).

TARDEC has taken an incremental approach to model development, beginning with a model of the "front-end" automatic, pre-cognitive aspects of human vision. This led to the TARDEC Visual Perception Model (VPM) of front-end visual processing and signal detection. The baseline VPM has been calibrated and validated as a predictive model of human performance in simple target detection tasks in automotive and military contexts [Witus 1996]. The baseline VPM represents automatic front-end visual processes including temporal filtering, color vision, multi-resolution spatial filtering, and nonlinear neural receptive field response in the striate cortex. The baseline VPM uses a highly simplistic "back-end" model to predict target detection, and does not represent performance of cognitive processes in target discrimination.

I.1.2 Objectives and Scope

The objective of this SBIR topic is to extend the baseline TARDEC VPM to predict human performance in visual target discrimination. The approach includes: (1) developing, testing, and integrating new algorithms into the model; (2) calibrating and validating the model in realistic perception

test scenarios; and (3) refining the model, then iterating the process. Separate but parallel trial design applications are being conducted in coordination with current materiel development programs and with potential downstream military and commercial users to ensure suitability of the VPM extensions for military and commercial applications of priority interest.

The objective of Phase I was to outline the technical approach for extending the baseline VPM for modeling visual target discrimination. The specific sub-objectives were to analyze the task and application requirements, formulate the modeling approach, demonstrate key algorithms, identify technical research issues, and develop a plan for Phase II. The objective of Phase II is to implement, test, calibrate, and refine the approach to produce a calibrated model of visual target discrimination.

By the end of the Phase II project we will have developed a predictive model of human visual discrimination. This model must balance the potentially conflicting requirements of simplicity, fidelity, general applicability to a wide variety of visual discrimination tasks, and specific applicability to priority applications. In fact, there is no one specific model realization that meets all of these conflicting requirements. Instead, the more promising approach is to develop a family of models which employ common modules representing elements of visual processing, and which employ a common architecture and modeling approach. Specific applications and contexts may then differ in the model realizations they employ for a given study. For example, camouflage evaluation applications may require a refined texture discrimination module, while automotive turn-signal discrimination may not require texture discrimination but may require a refined luminance adaptation module.

I.2 Summary of Results

The Phase I project identified shortcomings in the baseline VPM which limit its usefulness as a front end for target discrimination modeling. These shortcomings include luminance adaptation, texture perception, neural receptive field saturation, and edge response. Equations to correct these deficiencies are presented in Section II. Some of these equations were implemented in code as part of the demonstration of visual discrimination processing algorithms.

The Phase I project completed an analysis of the cognitive processes and types of prior knowledge involved in visual target discrimination. The analysis identified different processing modes for target discrimination. The processing mode active in any given time depends on a variety of factors including the quality of the signature, the target discrimination categories, and the expertise of the observer. We outlined the cognitive processing paths to model the different modes. These results are described in Section II.

The Phase I project developed flow charts and equations describing the approaches to modeling visual discrimination and candidate algorithms. These results are the basis for implementation in Phase II. The flow charts describing the approach to model the different visual discrimination modes, and alternative algorithms and equations for the component subprocesses were developed and are presented in Section II. The core subprocesses are image segmentation, multi-channel pooling, feature detection spatial filtering, spatial/logical induction and object categorization.

The Phase I project implemented and demonstrated key algorithms and equations for the component subprocesses and for selected upgrades to the baseline VPM. These algorithms and equations perform the following image processing functions: (1) generating a reference image by replacing the target with synthetic content which locally matches the surrounding texture and contrast, (2) computing the contributions of intensity and texture gradients to perception of the target shape, and (3) segmenting the image into object regions based on texture and contrast. These algorithms were demonstrated using a complex and highly heterogeneous scene with many different types of objects, textures and spatial relationships. The image processing illustrations are presented and described in detail in Section II.

The Phase I project outlined the research, test, and evaluation approach for Phase II. This approach focused on methods for iterative testing during development to ensure robust model capabilities and to develop calibration data for priority operational tasks. The Phase II approach is described in Section II, and includes illustration of methods to develop perception test stimuli by image transformation and deformation.

I.3 Conclusions

The Phase I project has successfully accomplished the main objective of outlining the technical formulation for modeling visual target discrimination and specifying details of the modeling approach. We implemented key elements of the technical model formulation as extensions to the baseline TARDEC Visual Perception Model at the TARDEC facilities, and ran the extensions to illustrate the analytic methods. These illustration results demonstrate the technical merit of the approach. We have also outlined and illustrated key elements of the approach for iterative model test and refinement in Phase II. Throughout Phase I, we have coordinated with potential downstream users in order to ensure that the modeling focus and emphasis addresses the dual-use military and commercial needs of TARDEC and industry.

The Phase I results provide a solid foundation for implementation, test, and refinement in Phase II. They provide a high degree of confidence in the technical merit and commercial value of the Phase II products.

SECTION II

This section documents the results of the Phase I project. There were five major technical tasks: (1) assessing critical shortcomings in the baseline VPM front end and designing upgrades; (2) performing task analysis of component processes and alternative modes of visual discrimination; (3) developing flow charts and equations describing the approaches to modeling visual discrimination and candidate algorithms; (4) implementing and demonstrating key algorithms; and (5) outlining the research, test, and evaluation approach for Phase II. This section documents the results of each of these tasks.

II.1 Upgrades to the Baseline VPM

Our technical coordination with potential downstream military ground vehicle and commercial automotive applications revealed that there are shortcomings in the current TARDEC VPM front end which need to be corrected before it will provide the appropriate outputs to the new target-discrimination back end being developed under this project. These upgrades are particularly relevant for camouflaged and concealed targets for which figure-ground discrimination is nontrivial, and for low-light conditions. Upgrading the retinal-cortical processing model (the VPM front end) is needed to provide adequate input to the visual cognition back-end models under development in the current SBIR project. Some of these shortcomings were also noted in the VPM documentation [Witus 1996].

The current VPM front end does not represent luminance adaptation. It does not represent texture-gradient perception or the contribution of texture gradients to object segmentation. It does not fully represent the sensitivity of the visual system to edges. It does not represent saturation in the neural receptive field (RF) nonlinear response module. These visual phenomena need to be represented in the VPM front end for accurate modeling of object segmentation and shape discrimination, especially for military targets employing camouflage, concealment, and deception. In addition to the visual processing issues, we identified needed enhancements in the VPM enabling software workbench used to implement the VPM. These enhancements are described in the following sections.

II.1.1 Nonlinear Luminance Adaptation Module

Luminance adaptation is the combined effect of pupil response, photo-pigment bleaching and regeneration, and neuro-electrical coupling. The transfer function is closely approximated by the following equation:

$$G = [\, 1 / (Y + \alpha * A^{\beta} + \gamma) \,] \qquad (1)$$

where G is the gain, Y is the luminance input image, A is the adaptation level, and α, β, and γ are constants. A module is needed to implement this equation. The module will compute a gain map, i.e., a different gain at each point in the image, based on the Y (luminance) plane, the luminance adaptation level, and the visual system parameter. The gain will be applied to each of the image planes corresponding to the S, M, and L cones before the linear luminance/color-opponent transform.

II.1.2 Oriented Receptive Field (RF) Spatial Filtering Module

Neural RFs act as two-dimensional spatial band-pass filters. Spatial filtering on the red-green and yellow-blue color-opponent channels is adequately represented by filters which have a circular annulus shape, i.e., band-pass in all orientations. On the luminance channel, most RFs are oriented (i.e.,

low-pass in one direction and band-pass in the orthogonal direction) with a median orientation bandwidth of 45 degrees. This processing can be represented by applying four convolution kernels (oriented at 0, 45, 90, and 135 degrees) to each multi-resolution plane on the luminance channel.

II.1.3 Between-Band Cross-Covariance Module

The covariance between the output on adjacent spatial frequency bands measures the correlation in phase and magnitude. The outputs on adjacent bands are in phase at edges, and the covariance analysis is needed to improve the sensitivity to edges and boundaries. The module to compute the between-band cross-covariance images will simply multiply the output from spatial filters at adjacent spatial frequencies and the same orientation. This will be done at the same point in the model that the in-band auto-covariance is currently computed (by squaring the outputs of the spatial filtering operations). This module should be connected in parallel with the current pyramid-squaring module in the nonlinear RF transfer function.

II.1.4 Second-Stage Texture-Gradient Response Module

The current VPM front end represents neural RFs that detect contrast gradients (luminance and color differences), but not texture gradients. Texture gradient response is essential for analysis of camouflaged vehicles. This module will compute the energy envelope (magnitude squared) of the output from the spatial band-pass filters, and then process this output with a duplicate of the first stage spatial filtering and nonlinear RF response module. This module will be applied only on the luminance channel, enabling the VPM to represent detection of boundaries by perceiving texture gradients.

II.1.5 Neural Receptive Field Saturation Module

The current equation for neural RF response is linear in the square of the contrast modulation, and does not represent the saturation nonlinearity. The current formulation is:

$$\text{RF\_Output} = CR_f^2 / (\alpha * CT_f^2 + \text{Boxcar}_n(CR_f^2)) \qquad (2)$$

A common analytic form to model saturation is:

$$\text{RF\_Output} = CR_f^2 / (CR_f^2 + \alpha * CT_f^2 + \beta * \text{Boxcar}_n(CR_f^2)) \qquad (3)$$

where CR_f denotes the contrast ratio image at a spatial frequency f, CT_f denotes the standard laboratory contrast threshold at spatial frequency f, α and β are empirical constants, and $\text{Boxcar}_n(.)$ is a spatial filter with a uniform square convolution kernel of width n. An alternative form is the negative exponential.

II.1.6 VPM Enabling Software Workbench

The VPM is implemented using a visual-processing-model development workbench. The workbench separates low-level C++ code from the high-level model flow. Individual modules become part of a library for use in different applications. This software implementation approach significantly reduced the time and cost of model implementation and testing. However, there are several limitations to the workbench which need to be corrected. It lacks automatic memory management. It has a clumsy

interface for displaying images and denoting the target regions of interest. The workbench needs to be implemented in two versions: a full-up developer's version, and a streamlined "execution-only" version for finished applications. These steps will reduce model development time and improve execution. The VPM lacks a graphic user interface for displaying and editing the visual-perception-model data flow diagrams. This interface is needed to reduce the time and cost of model development, tailoring, and documentation.

II.2 Task Analysis

The task analysis was based on a review of the relevant literature in computational cognitive psychology on visual discrimination, coordination with potential downstream military and commercial customers to identify application contexts and priorities, and personal experience in target acquisition with baseline and low-signature targets. We identified several distinct modes of target shape discrimination. Kosslyn's [1994] *Image and Brain* is devoted to an analysis of the mechanisms of visual recognition and identification. It is based on extensive empirical results of careful psychophysical testing with normal and brain-damaged human subjects, and neurophysiological testing with both human subjects and laboratory animals. The model of multiple paths to object categorization shown in figure 1 is based largely on Kosslyn's analysis, but is also consistent with personal experience in target acquisition testing, anecdotal evidence from military survivability technologists, and photo-intelligence analysis methods. Kosslyn's model, briefly summarized in appendix C , is too broad in scope and at too high a level of aggregation to be directly applicable. However the analysis underlying Kosslyn's model was a valuable resource for this project.

Figure 1 illustrates the alternative paths to target categorization. The different paths employ common modules, connected in different sequences. All begin with front-end retinal-cortical processing. One path goes immediately to feature detection spatial filtering by neurons tuned to specific shapes. This path is active when the signature is relatively high-quality and is clearly a separate and distinct object (i.e., not obscured and not connected to other objects), and the target has a simple and familiar distinctive shape, or distinctive component shape elements. The other path involves image segmentation and multi-channel pooling prior to feature detection spatial filtering. This path is active when the signature is relatively poor quality so that no individual color/luminance or texture channel has sufficient information to perceive the target as a whole and categorize it, when the signature is obscured, or when the target has a complex or unfamiliar shape.

II.2.1 Trivial Discrimination

Feature detection spatial filtering effectively compares the internal pattern of neural activation to the pattern corresponding to an ideal or iconic target or to characteristic components or features. When the characteristic shape is highly familiar, the overall target shape is the basis for spatial filtering (e.g., researchers have found individual neurons tuned to respond to the images of faces). In these cases the prior knowledge can be represented by the ideal or iconic form of the target.

In some cases, the overall target shape is complex and insufficiently familiar for immediate full shape recognition, but the target is clearly a distinct object and has its distinctive features or shape elements. In these cases, the prior knowledge for feature detection spatial filtering are the distinctive features or shape elements. In both of these cases the visual processing is essentially a visual pattern matching of the perceived signal to the prior mental image of the iconic or ideal form. Wolfe and Bennett

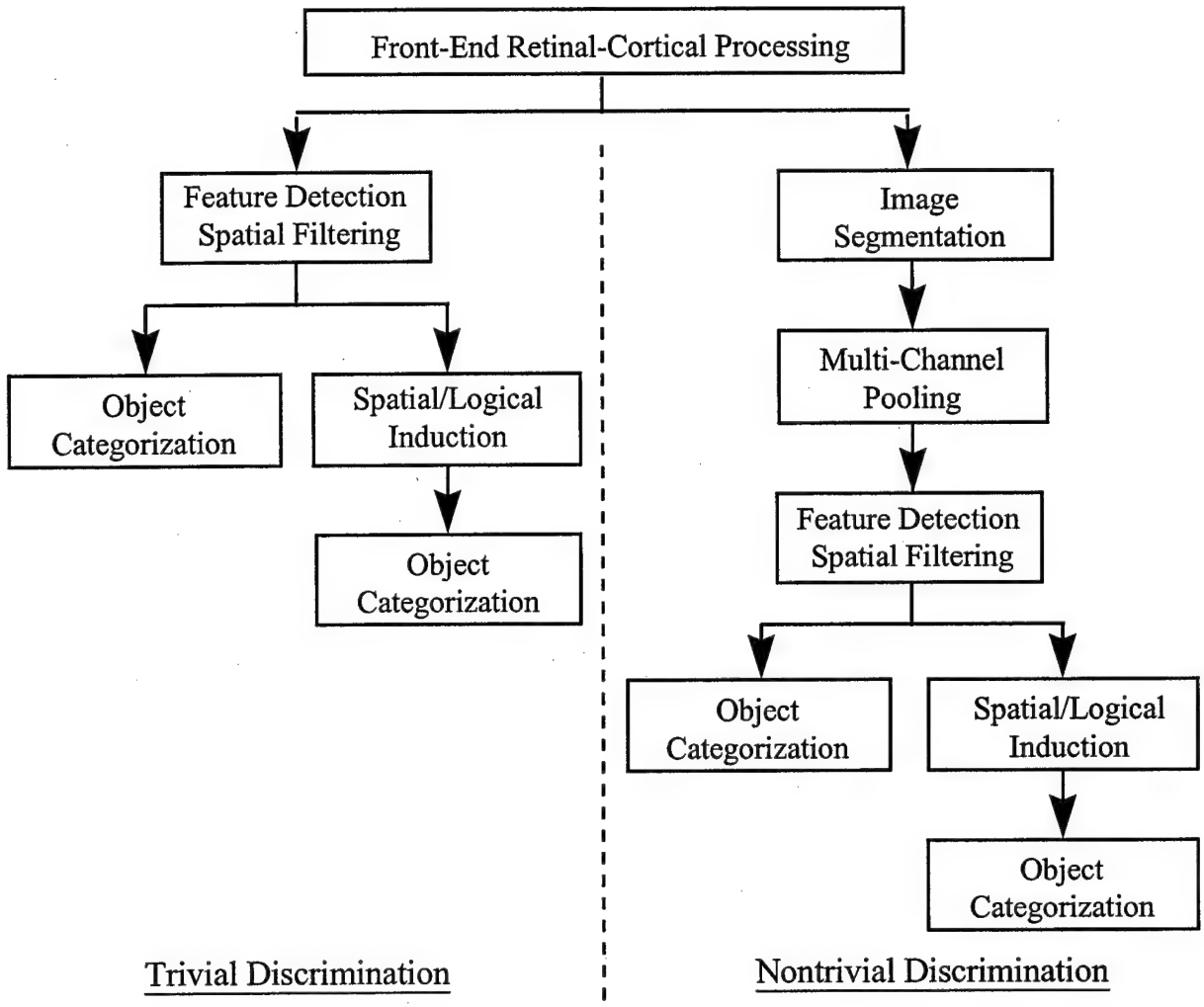

Fig. 1. Multiple paths for object categorization

[1997] present compelling empirical evidence that this is part of the automatic, massively parallel visual processing system, and is clearly distinct from serial processing when the target does not have simple distinctive features, or is not clearly separate from other objects.

When the target is clearly segmented from the surround, i.e., there is no obstruction or other adjacent objects to confuse the figure-ground discrimination, and when the target has a simple and distinctive overall shape or features, then the target discrimination is a parallel process and we have a "pop-out" target. When the target shape is complex, and lacking in distinctive features, but is still unambiguously segmented as a distinct object, then target is recognized based on the spatial and logical relations among the component features [Wolfe and Bennett 1995] [Kosslyn 1994: 54-104]. This *spatial/logical induction* requires serial inspection, and the target is no longer a "pop-out." In these cases the prior knowledge is represented by the visual appearance of the component shapes and the categorical relations or predicate calculus of the spatial and logical relationships among the component parts.

When the figure-ground discrimination is trivial, the visual cognition processing focuses on the individual features, or spatial/logical combinations of features, which distinguish one type of target from another. The signature of the target can be reduced to its distinguishing features.

II.2.2 Nontrivial Discrimination

In many real-world situations, the figure-ground discrimination is not trivial. The target may be partially obscured so that its parts appear unconnected, or the different parts may have significant color, luminance or texture gradients between them, or the target may have insignificant color, luminance or texture gradients between it and the objects adjacent to it. In these cases, figure-ground discrimination is required to perceive the target as an object, and this must precede recognition/identification. Even if the target has simple distinctive features or overall shape, it will not be a "pop-out" target because it does not have "pop-out" figure-ground discrimination. In these cases, figure-ground discrimination precedes feature detection spatial filtering. Figure-ground discrimination is performed in two stages: independent, single-channel image segmentation, followed by multi-channel pooling.

Image segmentation separates regions from surround. On each visual channel, the interior of a segmented region has similar value, and the surround has different values. There is a gradient between the region and its surround. Segmentation operates independently on the different color-opponent/luminance, temporal, and spatial band-pass channels. The segments do not necessarily correspond to physical objects. An object, especially complex objects built up from separate parts, may contain regions of different color, luminance or texture which cause it to be segmented into several regions. Shadows or other large-scale scene features (e.g., contrast with the sky) may create large segments which encompass several distinct objects.

Multi-channel pooling follows image segmentation. Split-and-merge (intersection and union) operations refine and consolidate the results of independent channel segmentation. Multi-channel pooling feeds feature-detection spatial filtering. When the figure-ground discrimination is nontrivial, then the visual cognition processing becomes a constructive process. The iconic features for feature detection spatial filtering are no longer the features which distinguish one type of target from another, but are the features or components from which the target signature can be constructed.

II.2.3 Object Categorization

Object categorization refers to the final step of deciding on the response: whether or not to assign the target to one (or more) of the prior target categories, and the degree of confidence in the categorization. There are two modes of decision-making in assigning a target to a particular prior category. These modes depend on the nature of the task. In multiple-alternative forced-choice (MAFC) tasks, the observer knows that a target is present and must assign it to exactly one category out of a set of mutually exclusive alternatives. In multiple-alternative open-choice (MAOC) tasks, the observer assigns a confidence rating to the different possible categories. A target may "possibly" belong to several different categories. Furthermore, in MAOC tasks the categories may not be mutually exclusive. These two modes of decision-making require different models. The MAFC case is easier, the common model is Baysean classification with a "winner take all" end-rule. [Kosslyn 1994: 119; Wandell 1995: 426-9]. The MAFC task is closely related to the detection of distinguishing features. In the MAOC task, we must also address whether there is enough information in the signature to make any decision at all. The MAFC paradigm is applicable when figure-ground discrimination is trivial and the target is not obstructed and not presented in an unusual perspective or configuration. In all other cases, the MAOC paradigm is more applicable.

Simple and highly familiar shapes are directly detected on individual neural receptive fields of the front-end visual system tuned to the particular shapes. This mode applies to "pop-out" targets: simple shapes, well trained observers, clean signatures, benign backgrounds, and extremely coarse discrimination selectivity. Ophthalmologists, neurologists, developmental psychologists, and vision

research psychologists may be interested in modeling the performance of visually or cognitively impaired individuals in response to "pop-out" targets.

Predicting observer performance for "pop-out" targets is not a focus of interest for the downstream military and commercial applications. The downstream application contexts are characterized by observers with normal vision in degraded viewing or low signature conditions and one or more of the characteristics of "pop-out" are missing: either the shapes are complex, the observers do not know what configuration the target will be in or how it will be obscured, the signature is suppressed and camouflaged, visibility is poor, the observer is not looking directly at the target, or the decision task involves target identification or recognition of subtle detail.

More complex and less highly familiar shapes, finer discrimination tasks, or targets with degraded signatures or cluttered backgrounds, are recognized only after segregating the target from its surround, integrating the multi-channel outputs of the front-end visual system's basic neural receptive fields, then analyzing the combined output to determine the shape of the target. The target must first be segmented from the background to be perceived as an object before the shape of the object can be ascertained. Portions of the target may be visible due to luminance contrast, other portions due to color or texture contrast, but the combined effects are needed to see a sufficient portion of the target to recognize it. Camouflage and signature management often concentrate on defeating the figure-ground discrimination process. If it is not perceived as an object, it will not be recognized as a target.

Both of these modes apply to relatively simple targets consisting of a single region in which the shape of the target is characterized by its outer boundary. The visual resolution of the boundary is sufficient to measure the information available to discriminate the target, and target categorization is a matter of matching the target outline to the iconic forms characteristic of the different categories. The measure of perceived information on the boundary is useful to predict how well people are able to discriminate the target. The measure of relative similarity of the perceived signature to that of the icons or ideal images characteristic of the different target categories is useful for predicting what target categorization decisions people will make. These are two different prediction problems, which require different model formulations.

Highly complex objects which consist of a spatial organization of a number of components involve recognizing individual components in a spatial and/or logical configuration characteristic of the complex object. This mode requires first categorizing the components of the target, then matching the spatial and/or logical relationships among the components to templates for alternative types of targets. This mode of discrimination applies to very complex targets, fine levels of target identification, images with poor visual signature, or images so cluttered that even simple or familiar shapes cannot be recognized directly. This is particularly important when the target may be presented at noncanonical orientations or when its components are in different relative orientations. Kosslyn [1994: 241] concludes that "one identifies contorted objects by recognizing individual parts and distinguishing characteristics in the pattern activation subsystem, and computing the categorical spatial relations among them in the dorsal system."

In this situation the feature detection spatial filtering icons do not correspond to the entire target, but only to characteristic components or shape elements. Some features may be important for constructing the target (they are important for the observer to have confidence that he is looking at a target), but may not be valuable for distinguishing between alternative categories because the features are common to several categories. In particular, some features may be common to a broad class of targets, e.g., military combat vehicles, and distinguish that class of targets from other classes of targets. But these features will not be of value for distinguishing types of targets within the broad class precisely because they are shared features. Other features may not contribute significantly to confidence that the object is a target, but may be important for distinguishing between target types. This mode of discrimination does not necessarily imply conscious, deliberate inspection (although this may occur), but it does imply use of

categorical knowledge of spatial and logical relations among parts rather than strictly visual image matching.

II.3 Approach to Modeling Visual Discrimination

Visual discrimination modeling is focused on situations with degraded signature or viewing conditions (the nontrivial paths in figure 1), and excludes the more direct discrimination paths for "pop-out" targets (the trivial paths in figure 1). The processing flow for "pop-out" targets involves a subset of the processing functions for degraded scenes. Should modeling discrimination performance for "pop-out" targets become a priority, the models developed for degraded scenes should be able to be reconfigured for "pop-out" targets.

The proposed visual discrimination modeling approach and alternatives are illustrated in figure 2. In section II.3.1 we have outlined two different approaches to modeling performance in visual discrimination subprocesses (image segmentation, multi-channel pooling, and feature detection spatial filtering). In section II.3.2 we describe two alternative approaches to modeling spatial-logical induction processing. In section II.3.3 we describe formulations to predict object categorization performance for MAFC and MAOC tasks.

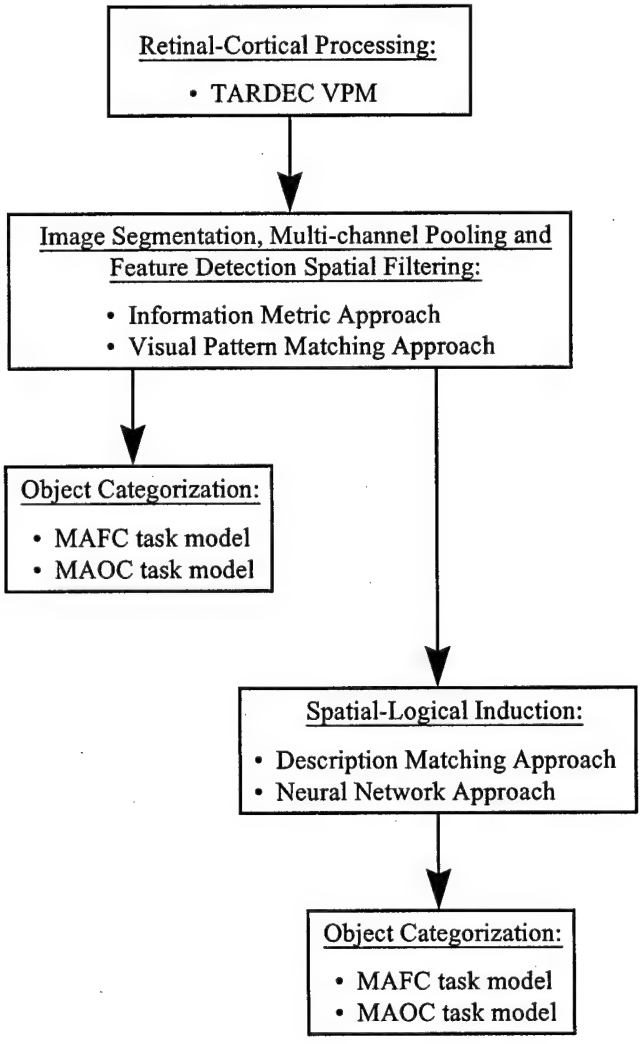

Fig. 2. Nontrivial Visual Discrimination Modeling Approach and Alternatives

II.3.1 Modeling Image Segmentation, Multi-channel Pooling, and Feature Detection Spatial Filtering

There are two alternative approaches to modeling image segmentation, multi-channel pooling, and feature detection spatial filtering: (1) an information metric approach, and (2) a visual pattern matching approach. The information metric approach attempts to predict performance from a measure of the useable information in the image. It is an evolutionary modeling approach in that it is a generalization of the signal detection framework of the baseline VPM. The visual pattern matching approach is new in that it attempts to explicitly represent the processing of vision subsystems.

The information metric approach attempts to predict observer performance based on a measure of the amount of information the visual system can extract from the image, relative to the amount of information needed to perform the discrimination task given the shape of the target object. This approach does not predict what the observer will do with the information. Suppose a tank were constructed so that it looked like a cow. This model would compute how much visual information there was to categorize the shape, relative to the amount of visual information needed to categorize the shape. It might predict that the subject had sufficient information to categorize the shape, but it would not predict whether the subject would report seeing a cow or a tank. This modeling approach is promising for camouflaged and concealed targets, but not for deception. The information metric approach does not attempt to mimic the internal cognitive processing. It is an evolutionary approach because it is an extension of the psychometric modeling approach used for target detection in the baseline VPM. This approach does not explicitly represent the internal mental visual models of target appearance.

The visual pattern matching approach attempts to mimic the internal processes of mental imagery and image comparison. The visual pattern matching approach is model-based, in that it attempts to match the internal visual representation to a prior model. It is a much more ambitious, and hence risky, approach in terms of technical feasibility, time and financial resources. Well-respected researchers (e.g., Biederman [1987], Lowe [1985, 1987a, 1987b]) have pursued this approach with only limited success, and then only for line drawings and low-noise easily segmented images. Variations on this approach, but without the multi-channel front end, have been proposed for Automatic Target Recognition, and have also met with only limited success. Applying the method is also difficult because it requires external specification of the iconic target forms, which are a function of observer training, operational task, anddiscrimination mode (which depends on image quality). However this is the only approach which has potential to adequately handle decoys and deception.

Both approaches are applicable to simple targets which do not require spatial/logical induction, and to complex targets which do. At this time it appears that a common approach to spatial/logical induction can be developed for use with both the information metric and visual pattern matching approaches. Both approaches are applicable to MAFC and MAOC object categorization tasks. The information metric approach predicts the probability of correct response. The visual pattern matching approach computes the probability of each response.

Section II.3.1.1 describes the information metric approach for situations in which target signatures do not require spatial/logical induction for target discrimination. Section II.3.1.2 describes the visual pattern matching approach, also for situations in which target signatures do not require spatial/logical induction for target discrimination.

II.3.1.1 Information Metric Modeling Approach

This approach computes the perceptible information available to the observer for discriminating the target from its surroundings and to determine the shape of the target. Kosslyn [1994: 226] finds that:

> Two sorts of top-down control strategies appear to underlie the patterns of eye movements that occur when people cannot identify an object at a glance. First, in some cases, top-down attentional control is driven by a specific hypothesis, such as that one is viewing a cat and hence should look for whiskers at the front of its face. Second, in some cases, subjects do not appear to be testing a specific hypothesis, but instead engage in systematic search strategies. . . . If the input is weakly consistent with many possible objects, a good strategy is to scan the object systematically looking for more information. The highest-information parts of an object are often along its contour, and hence top-down mechanisms might simply lead to scan along the object's boundary.

The information metric model measures the information available from the target boundary, whether the information is perceived at a glance or by scanning along the boundary. It measures the information available for making a decision relative to the information needed to make a decision. Image segmentation is implicit in the information metric modeling approach in that an analyst must outline the region of interest, much like the current VPM. Multi-channel pooling is explicit: the probability that a spatial unit of information on the boundary (at each multi-resolution scale) is resolved is a function of the pooled signal over all channels at that position and spatial frequency band. This approach computes the number of units of resolvable spatial information integrated over the boundary. It also computes the number of units of information needed to recognize the target shape as a function of a set of shape complexity factors. The psychometric function to predict subject response compares the amount information available to the amount of information required.

The flow for the boundary shape information metric analysis is illustrated in figure 3. The RGB image is first converted to the internal luminance/color-opponent coordinate system. The next step is to analyze the texture magnitude on the luminance channel. The luminance, color-opponent, and texture planes are individually analyzed to compute its neural RF response image.

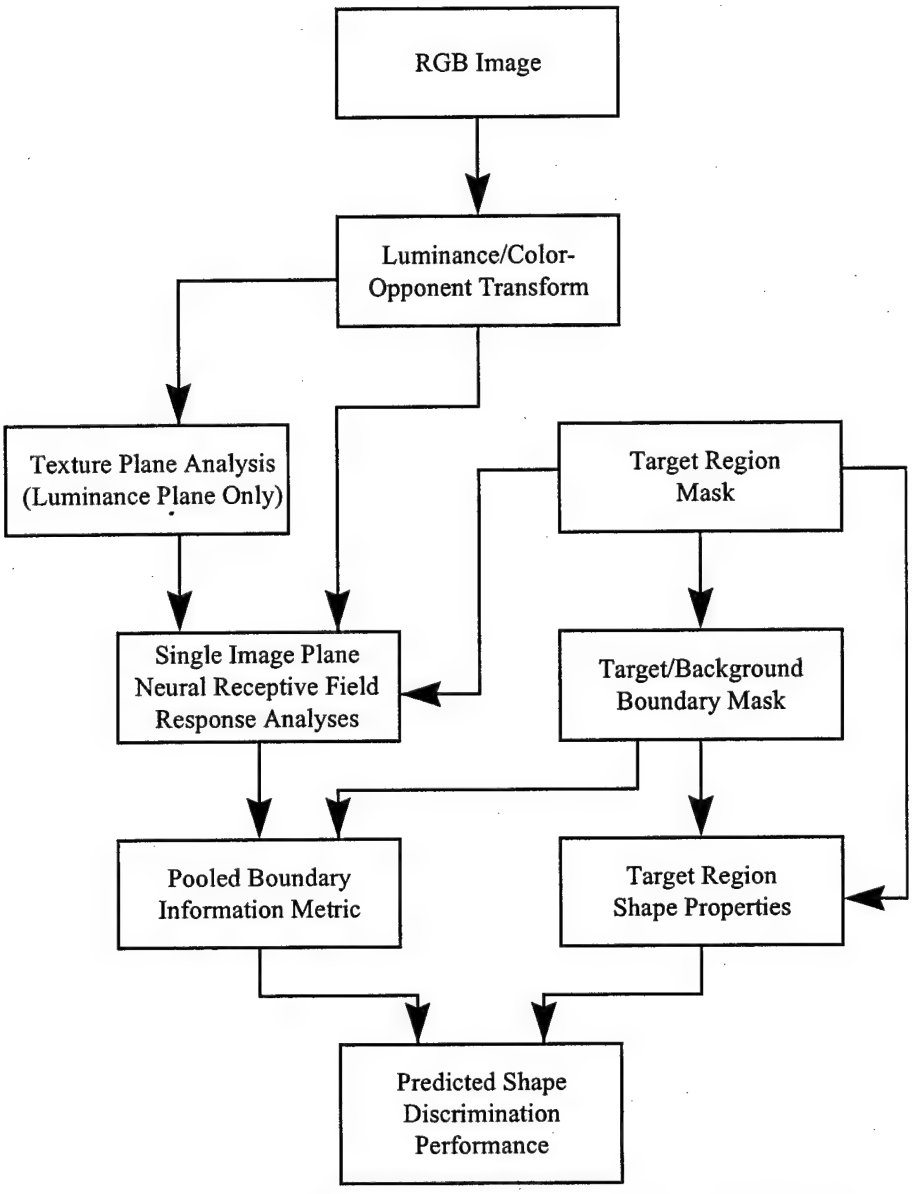

Fig. 3. Overall Model Flow for Boundary Information Metric Model

Figure 4 illustrates the single image plane neural RF response analysis. This is nothing more than an extension to the baseline VPM to include luminance and texture channels. The background bias image is a key element in this processing. It represents the visual content due to the background, independent of the target and without target-background interactions. The background bias image represents "zero." The neural RF response to the background bias image is subtracted from overall neural response in order to determine what response is due to the target. The spatial filtering module is essentially the current VPM spatial filtering, with the addition of orientation filtering on the luminance channel. The normalization and adaptive gain formulations are essentially the current VPM formulation with the addition of RF saturation.

Next the neural RF response maps are pooled over the luminance, texture, and color-opponent channels. The initial pooling formulation is power-law summation [Graham 1989: 164-80]. It is simply the nth root of the sum of the nth power of the RF output on each channel. This is computed for

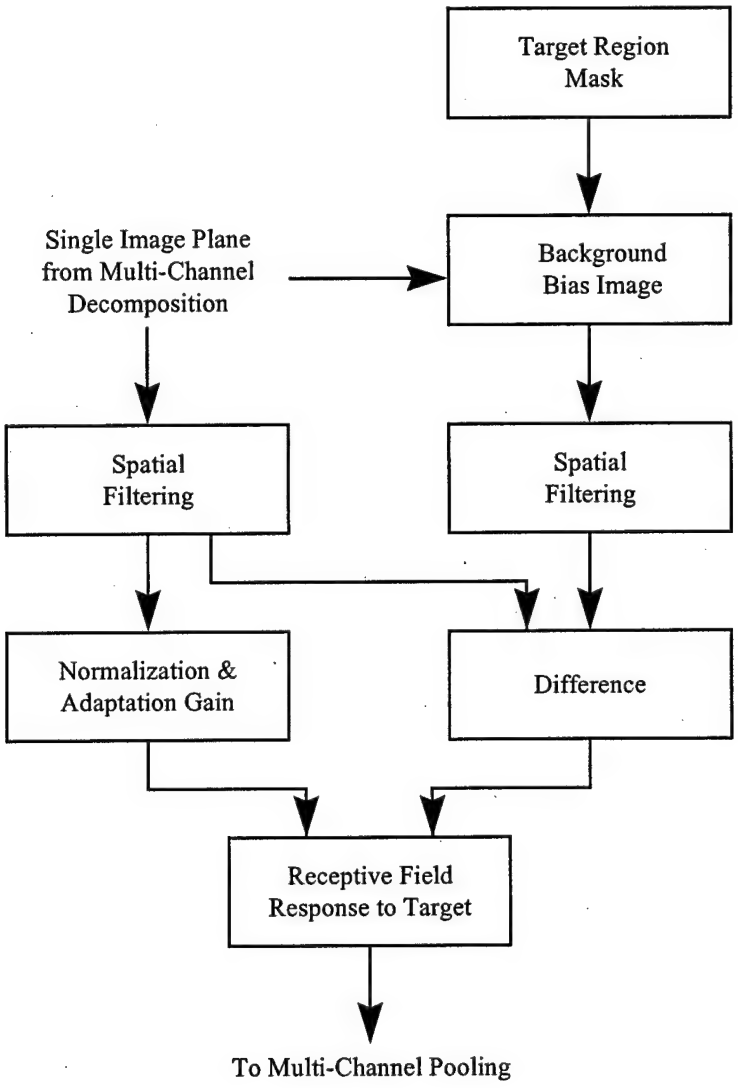

Fig. 4. Single image plane neural receptive field response model

each multi-resolution pixel location (i,j) for each spatial frequency b and f, each orientation θ on the luminance channel, and each color-opponent/luminance/texture input channel:

$$\text{Pooled\_RF\_Response}_f(i,j) = [\ \Sigma_{cf\theta}\ \text{RF\_Response}_{cf\theta}(i,j)^n\ ]^{1/n} \qquad (4)$$

The pooling stage also includes a nonlinear transformation to yield the probability of detecting each RF location in the image. The initial analytic form for the cumulative distribution function is the negative exponential distribution. There are two calibration parameters: the input level for 50 percent probability of detection, P_{50}, and an exponent, β, governing the spread of the distribution.

$$\text{Prob\_Detect}_f(i,j) = 1 - e\ xp(\ \ln(1/2) * (\ \text{Pooled\_RF\_Response}_f(i,j)/P_{50}\ )^{\beta}\ ) \qquad (5)$$

These probabilities are integrated over the target boundary to yield the expected number of detectable RF locations on the target boundary. This yields the boundary information vector. There is one component for each spatial frequency band. The cumulative information is simply the sum over the

spatial frequency bands. In essence it measures the number of units of spatial information the visual system detects on the target boundary.

This metric by itself is not sufficient to predict target discrimination and shape perception performance. Complex shapes require more information than simple shapes. We need to compute a measure or measures of the target shape complexity, then compute the shape discrimination performance from the shape complexity metric and the amount of visual shape information. A variety of shape metrics have been proposed in the literature [Brady and Yuille 1987: 329-60] [Gonzalez and Wintz 1987: 391-414]. In general, shapes with more concavities, more disjoint regions, more holes, more fine detail, more asymmetries appear more complex. We have found no basis to prefer any particular shape metrics in the literature on visual cognition, nor have we found any *a priori* reasons in the digital image processing literature to select any particular metrics as measures of complexity for visual cognition. The selection of the shape parameters, and the form of the psychometric function to predict discrimination performance from the boundary metric and shape complexity are empirical questions for Phase II. Some of the promising shape complexity metrics to be considered in Phase II include the following:

1. the area, perimeter, ratio of the area to the square of the perimeter (i.e., the compactness), length of the major axis, length of the minor axis, and ratio of the major axis to the minor axis;
2. the moments of the distribution of the distance from the points on the boundary to the center of mass of the boundary (or center of mass of the region), especially the mean and variance; and
3. the area and perimeter of the smallest circumscribing ellipse, the area and perimeter of the largest inscribed ellipse, and the ratios of their areas and perimeters to the area and perimeter of the region.

For whole target recognition, i.e., when spatial-logical inference is not applicable, the analyst designates the outline of the whole target. For complex targets, the analyst outlines and evaluates each of the target components individually to generate the input for spatial-logical inference.

II.3.1.2 Ideal Image Matching Modeling Approach

One of the main conclusions of Kosslyn [1994] is that visual pattern matching between the perceived image and a mental image of an ideal or iconic form (either an entire object or principal components) is a key element of visual cognition. Kosslyn also concludes that this visual pattern matching process includes rotating, re-scaling, and stretching to register the perceived object with the icon, before evaluating the strength of the match. Kosslyn further asserts that the icons are stored as images, and not reconstructed from a propositional description of the targets. He sites Standing's [1973] experimental results on human visual image recall, showing the vast capacity and high processing rate for mental imagery. Standing calculated that normal subjects had a mental search rate for specific pictures of 50,000 images per second, with a 99 percent accuracy for near-term recall, and 73 percent accuracy for delayed recall, with a short-term capacity for over a million pictures.

The objective of this model is to measure the strength of association between the perceived form of the image and the perception of an ideal target or iconic form characteristic of the alternative category choices, and from this to predict the probability that an observer will make a given categorization decision. The technical formulation is very similar to that previously described for measuring the perceptual boundary information. The major difference is that instead of integrating the perceptual boundary information over the target boundary, we compute the correlation between the perceptual boundary information from the image and the perceptual boundary information from an ideal target or iconic form.

Bergen and Landy [1991] used this approach as the back end to a simple computational vision model front end to predict successfully observers' ability to discriminate the orientation of asymmetrical

targets. This approach, using an ideal target and ideal observer as a reference point to evaluate the quality of an image, is described in detail in Geisler [1989].

The ideal target is a high contrast target image in a low noise background, with the minimal size for performance of the discrimination task [Nakayama 1990]. The modeling steps are: (1) process the actual image to obtain the neural RF output, (2) process the ideal image to obtain the theoretical RF output, (3) re-sample the actual image to scale the target to the same size as the icon, then (4) compute the correlation of the RF responses for the actual and ideal targets.

This process compares the target to each of the ideal targets representing the different target categories for whole target matching, and to component categories for input to spatial-logical induction. Biederman [1987] attempted to demonstrate a single, universal set of iconic forms ("geons") underlying visual pattern matching (within the scope of his line-drawing stimuli). It is unlikely that such a set of universal forms exists. Instead the set of target categories and iconic forms is entirely task dependent, and is input to the model. The set of icons depends on what features distinguish the object categories, and what features are common to the object categories. Further research is needed to establish guidelines and procedures to define the response categories and corresponding iconic forms. We have formulated the following preliminary guidelines for further refinement in Phase II:

1. The categories must be narrowly enough defined so that instances of target masks are more highly correlated with members of their own category than they are with instances of other categories.
2. The icon for each category is essentially a mask, i.e., high-contrast, low-clutter image.
3. The resolution of the icon images should be the lowest resolution consistent with a designated threshold level of response, e.g., X percent correct discrimination for all icons in MAFC tasks, or for MAOC tasks, X percent correct discrimination at Y percent error at some specified confidence level.
4. Candidate iconic forms are: (a) the average of the target mask instances in the category and (b) the one instance with the highest correlation with all other instances in the category.

II.3.2 Modeling Spatial-Logical Induction

The spatial-logical induction module is used when the target type is inferred from the recognition of component parts in a characteristic logical or spatial relationship. Logical and structural models are needed to represent target discrimination through logical deduction or inference [Medin and Ross 1991: 128-34]. These models provide a framework for combining predictions regarding the discrimination of components to predict the discrimination of the whole. This type of approach to modeling very high-level aspects of visual cognition has been employed by Feldman [1987]. The computed inputs to the spatial-logical induction module are the probabilities that the component parts are correctly categorized (using the MAOC task formulation described in section II.3.3). These probabilities can be computed either by the information metric approach or the visual pattern matching approach. The module also requires inputs which specify what characteristic spatial and logical relations among the component parts characterize the different types of targets.

We have identified two alternative approaches to modeling spatial-logical induction: (1) description matching, and (2) neural network. Both approaches require that an analyst specify in advance the target categories, the types of components which characterize the targets, and the types of spatial-logical relationships between components. The description matching approach requires that the analyst define the specific spatial-logical relationships which characterize each target category, whereas the neural network approach requires that the analyst provide a robust set of exemplar training images which

characterize the target categories. These two approaches are described in sections II.3.II.1 and II.3.2.2, respectively. The description matching approach represents the spatial-logical relationships with a formal grammar, whereas the neural network approach uses the connection weights on the hidden layer.

II.3.2.1 Description Matching Approach

The description matching approach employs a generalized "And/Or" graph to represent the spatial and logical relationships among components of the target, and their contribution to overall target discrimination. In its simplest form, this approach assumes all leaf nodes are statistically independent and the graph has a true tree structure with alternating "And" and "Or" nodes traversing from root to leaf. The leaf nodes represent facts. Each fact has a probability of being true. The probability that the root node is true is evaluated recursively: the probability that an "Or" node is true is the probability that *any* of its branches are true; the probability that an "And" node is true is the probability that *all* of its branches are true. If all nodes are independent, then traditional probability calculus can be used to evaluate the "And" and "Or" nodes. The value of an "And" node is the product of the probabilities of its branches. The value of an "Or" node is one minus the product of one minus the probabilities of its branches.

In general, the nodes are not independent, and the same node may be on two different branches. In this case the graph is not a true tree structure and the fuzzy logic calculus of possibilities is used instead: the value of an "And" node is the minimum of the values of its branches, and the value of an "Or" node is the maximum of the value of its branches. Fuzzy logic approaches have shown significant potential for use in predictive models of target discrimination [Singh et al. 1996].

A generalization to this framework is to include a "weight" between zero and one for each node. The gain represents the confidence or value of the node. It is the maximum contribution the node can make, even when its possibility value is unity. This generalization has been effectively used in risk analysis [Schmucker 1984: 43-77]. The US Army Ballistics Research Laboratory (BRL) uses this approach to model the effects of component damage on vehicle function. In the vulnerability analysis, backup systems may be able to perform a function, but not as well as the primary system.

This weighted, fuzzy-logic "And/Or" graph is the basis for the logical induction model framework. However, it must be extended to include spatial relationship nodes in addition to the strictly logical "And" and "Or" nodes. Spatial relationships are binary nodes which are evaluated like "And" nodes if the spatial relationship is true, but has value zero if the spatial relationship is false. If the spatial relationship is true, its strength is the minimum strength of the two components. We also need to add a "Not" relationship, since the perception of components can be used to reject possibilities. The "Not" relationship is evaluated as one minus the input value. Kosslyn [1994: 192-214] addresses the transformation from coordinate to categorical spatial relations, and suggests a number of categorical spatial relationships in the internal encoding of object structure. These relationships include the following:

1. inside / outside
2. connected to / not connected to
3. above / below
4. left of / right of
5. close to / far from
6. similar size as / larger than
7. similar orientation / orthogonal relative orientation / 45-degree relative orientation.

Kosslyn does not propose quantitative definitions for the relations which categorically encode relative size or distance. We propose to explore using the number of orders of magnitude, base 2, for the coordinate-to-categorical transformation. So the categorical values of the relative size of two components would be the base 2 logarithm of the ratio of their size, rounded to the nearest integer. Relative distance (close/far) can be categorized relative to the size of the larger component and smaller component using the same transformation.

II.3.2.2 Neural Network Approach

We consider the use of neural nets to be highly exploratory. Most of the examples of neural nets as models of visual processing have been demonstrated only for relatively small problems and simple stimuli. In all cases that we encountered, they were used to actually categorize the stimuli, not to predict the probability of correct observer classification.

The success and robustness of any use of neural nets is highly dependent on the quality of the training image set. The selection of the image set requires that the analyst identify in advance the various possible spatial relationships among component parts, and select a set of images representing all of these combinations, with and without degraded appearance. Given that the analyst has to develop this knowledge, he could then simply define the characteristic relations using the extended spatial And/Or graph structure. Since the neural nets represent interactions and nonlinear relationships, the set of training images must represent the full set of combinations, and design-of-experiment methods cannot be used to reduce the training set.

There are several alternative uses of neural nets. In the simplest approach, the inputs to the neural net are categorical spatial relations to the neural net, and the probability of categorizing the components. The neural net has only to compute the strength of association with the different categories. Another avenue of investigation is input images showing the strength of association of each component/feature at each different position. The neural net must then determine the spatial relationships and then the strength of association with different categories based on the probability of recognition of the components and their spatial positions in the image. Kosslyn et al. [1992] employed a neural net in this manner, but only for a very small problem (two types of components and four possible positions). Another avenue of investigation is to input the categorical spatial relations to the neural net, and the probability of categorizing the components. We are not proposing to use the neural net determine the nature and location of the characteristic components. Neural nets have been used this way, but only for extremely small artificial retinas, and extremely simple figures [Rueckl et at. 1989].

II.3.3 Modeling Object Categorization Performance

When the input to this stage comes directly from the information metric model, we can only evaluate the probability of correct response. We propose to calibrate a simple psychometric function to predict the probability of correct categorization, e.g.:

$$\text{Prob\_correct} = 1 - \exp\left(\ln(1/2) * \left(\left(\text{InformationMetric} - k \right) / f(\text{Shape Complexity}) \right)^r \right) \quad (6)$$

where f(Shape Complexity) expresses the value of the information metric for 50% probability of correct response, k is the response bias, and r is a steepness factor. The calibrations will be different for different sets of prior categories, and will be different for MAOC and MAFC tasks. It is possible that the response bias, k, and the function f may also need to be function of the correlations among the alternative categories. This is an empirical question that will be examined further in Phase II.

When the input to this stage comes from the visual pattern matching or spatial-logical induction, it will receive a strength of response, S_c, for each target category c. In MAOC tasks the subject has can

give a positive response to more than one category at any given confidence level, or can express a confidence level for each category. For MAOC tasks the probabilities of positive response do not need to sum to unity. For MAFC tasks they must sum to unity. Kosslyn [1994: 119] observes that each choice inhibits every other choice, and the strength of inhibition is proportional to the strength of activation, and that the recognition process depends both on the absolute and relative strength of activation. A candidate form for the probability of positive response in the MAOC task is

$$MAOC\_prob\_respond_c = (S_c / (S_c + k)^{(S_c/X)} \quad (7)$$

$$X = ((1/n) * \Sigma_i S_i^r)^{1/r} \quad (8)$$

where n is the number of categories, k and r are calibration parameters. The exponent (Sc/X) represents the inhibition effect. As r becomes large, the term X approaches the maximum value of S_i.

For MAFC tasks the candidate model is simply the MAOC probability of response, normalized to the sum of the probability of response over all categories:

$$MAFC\_prob\_respond_c = MAOC\_prob\_respond_c / \Sigma_i MAOC\_prob\_respond_i \quad (9)$$

II.4 Demonstrating Selected Key Algorithm Components

This section describes and illustrates selected key algorithm components implemented in Phase I. We demonstrated a technique to generate background bias images for detection of target boundaries due to first-stage luminance gradients and due to second-stage texture gradients. This was selected as a priority topic for early demonstration because of the critical role of the background bias image in model formulation and because "inverting the model" to generate the background bias image demonstrates the strengths and limitations of the formulation. In implementing the background bias image synthesis algorithm, we also implemented and demonstrated methods for evaluating texture and adding orientation filtering to the spatial processing.

We demonstrated the algorithms for single plane neural RF response analysis. This analysis produces the RF response images which are input to multi-channel pooling. These techniques are demonstrated on the luminance gradient and texture gradient channels using the luminance and texture background bias images. The demonstration showed how these analysis methods measure the perceptibility of the target boundary. The results are very significant because they demonstrate that this approach will produce meaningful target boundary information metrics, even in very complex and heterogeneous scenes.

We demonstrated algorithms for multi-resolution region segmentation, including algorithms to distinguish between modulation due to boundaries between objects and modulation due to the texture within an object. These results are significant because it is necessary to make this distinction in order to segment whole objects when the objects have internal texture (e.g., a type of foliage or camouflaged target), and not break the object up. However it is difficult because the appearance of texture is due to the boundaries between many smaller regions.

In order to perform these demonstration analyses, a number of new modules were added to the baseline TARDEC VPM. These modules are listed in Appendix B, and were installed at TARDEC for the demonstration analyses.

II.4.1 Background Bias Image

Figure 5 illustrates that the spatial pattern analysis model used in this demonstration has two

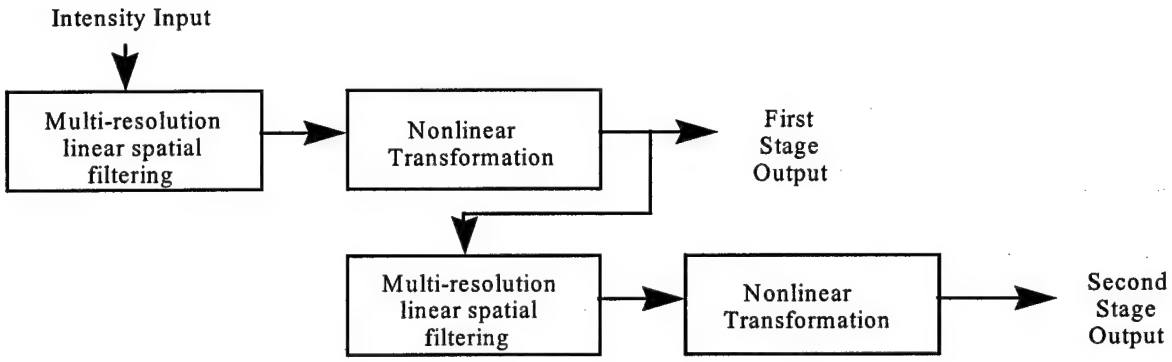

Fig. 5. Spatial filtering flow diagram

stages of spatial processing. Each stage consists of linear spatial filtering, followed by a nonlinear transformation. Multi-resolution linear spatial filtering is implemented by multi-resolution spatial band-pass filtering without orientation selection, followed by orientation filtering. The spatial filtering flow is illustrated in figure 6. The band-pass filtering is implemented as a difference of Gaussian low-pass filters:

$$L_0(\text{Input}) = \text{Initial Input} \tag{10}$$

$$L_{i+1}(\text{Input}) = \text{Subsample}(K * L_i(\text{Input})) \tag{11}$$

$$B_i(\text{Input}) = L_i(\text{Input}) - \text{Expand}(L_{i+1}(\text{Input})) \tag{12}$$

$L_i(.)$ denotes the i^{th} multi-resolution level of low-pass filtering; $B_i(.)$ denotes the i^{th} multi-resolution level of band-pass filtering; Subsample(.) denotes 2:1 horizontal and vertical subsampling; Expand(.) denotes the inverse of Subsample(.), i.e., 1:2 spacing with linear interpolation in each direction; and K denotes a convolution kernel discrete approximation to a 2-D Gausian filter. The multi-resolution process stops when the minimum dimension of $L_i(.)$ is equal to one. $L_{max}(.)$is the low-pass residual.

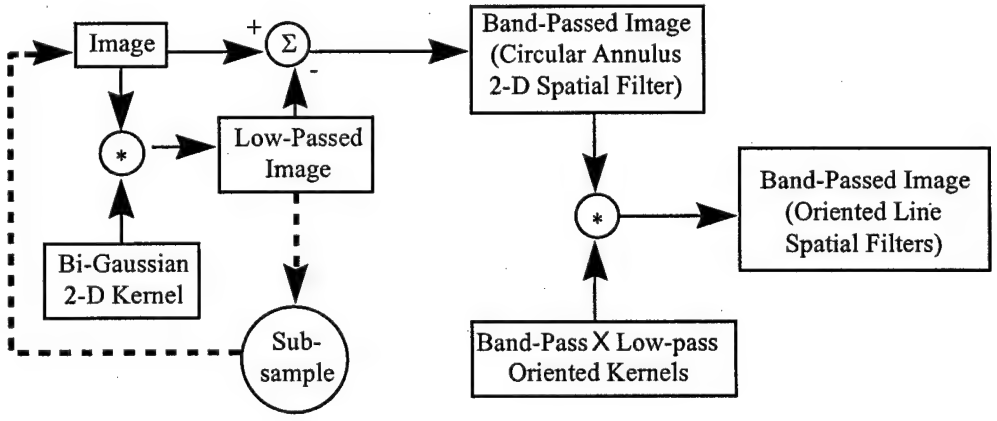

Fig. 6. Multi-resolution linear spatial filtering flow diagram

Each step of the spatial band-pass filtering is at one octave lower frequency. The subsampling process stores each low-passed image at the Nyquist limit for efficient storage. The result of multi-resolution spatial band-pass filtering is illustrated in figure 7.

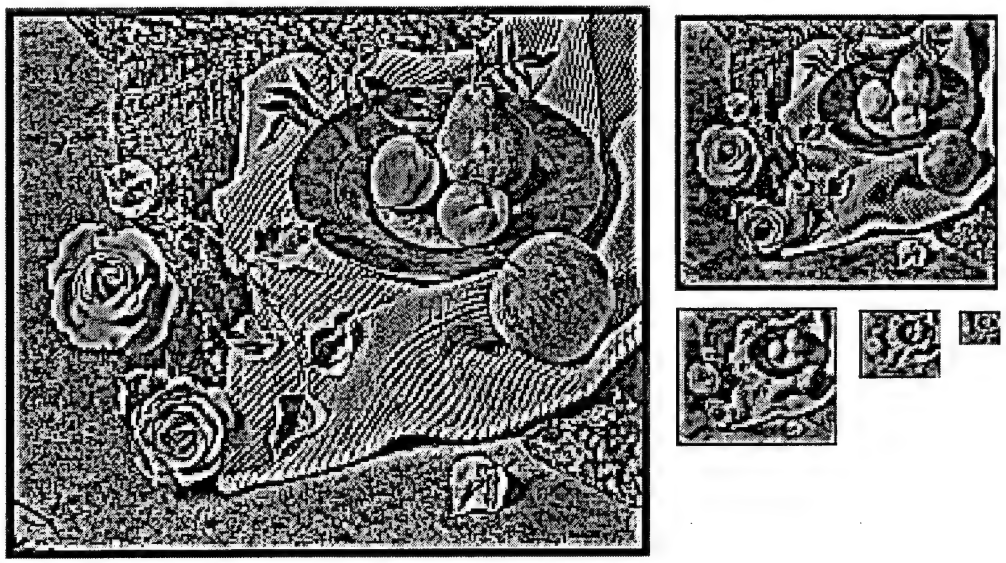

Fig. 7. Example multi-resolution band-pass pyramid

The inverse band-pass filtering operation simply involves expanding and summing the band-pass planes with the low-pass residual:

$$L_i(\text{Input}) = B_i(\text{Input}) + \text{Expand}(\, L_{i+1}\,(\text{Input})\,) \qquad (13)$$

$$\text{Reconstructed Image} = B_0(\text{Input}) \qquad (14)$$

The orientation filtering is implemented via convolution of the band-pass output $B_i(.)$ using kernels which are band-pass in one orientation and low-pass in the orthogonal direction:

$$C_{i\theta}(\text{Input}) = K_\theta * B_i(\text{Input}) \qquad (15)$$

K_θ are the orientation filtering kernels. In the trial demonstration, we used four orientation filtering kernels, with the low-pass orientations at 0, 45, 90 and 135 degrees. The kernels were chosen so that kernels representing orthogonal directions were orthogonal (i.e., their inner products were zero), so that the kernels sum to the identify transform, and so that the filtering in the low-pass direction is a discrete approximation to a Gaussian.

Table 1 shows the kernels for unoriented low-pass filtering and oriented band-pass X low-pass

Unoriented	K_0	K_{90}
1/16 1/8 1/16 1/8 1/4 1/8 1/16 1/8 1/16	0 -1/8 0 1/8 1/4 1/8 0 -1/8 0	-1/8 0 1/8 0 1/4 0 1/8 0 -1/8

Table 1. Kernels for Unoriented Filtering and Oriented Filtering at 0 and 45 Degrees

filtering at 0 and 45 degrees. Figure 8 illustrates the impulse response of the initial band-pass filter, and

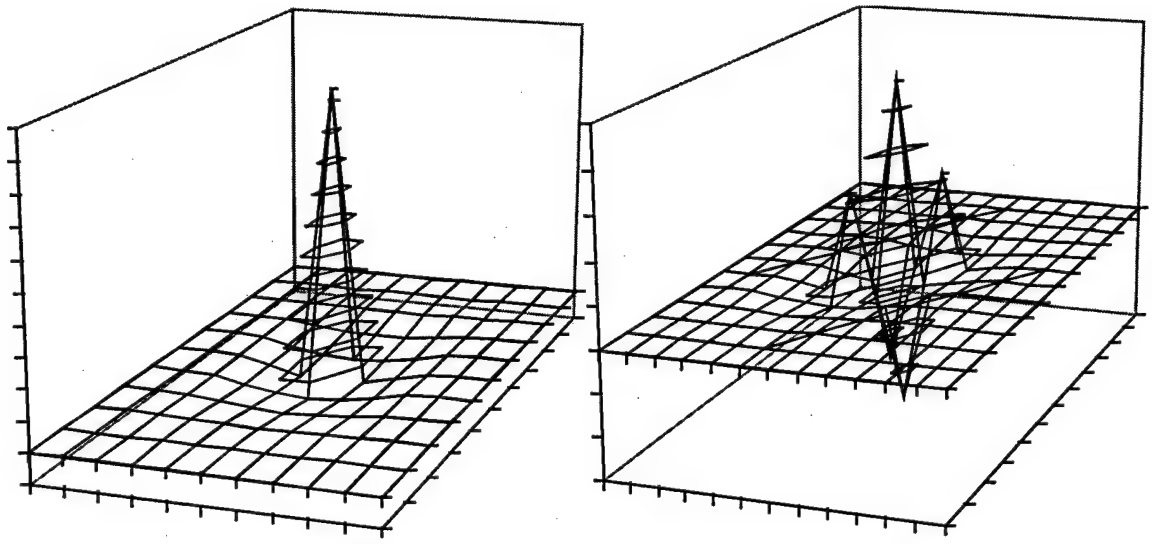

Fig. 8. Impulse response of 2-D band-pass filter, and 2-D band-pass filter followed by a 2-D band-pass X low-pass filter

the impulse response of the band-pass filter followed by the 45-degree-oriented band-pass X low-pass filters. With this choice of kernels, the inverse 2-D orientation filtering is simply summation over orientation:

$$D_i(\text{Input}) = \Sigma_\theta C_{i\theta}(\text{Input}) \qquad (16)$$

We chose four orientations at 45-degree intervals because 45 degrees is the median orientation bandwidth of human visual neural RFs. Actual RFs occur at all orientations with a distribution of orientation bandwidths. The computer model is a discrete approximation. We note that these four kernels do not constitute a basic set, but neither do the neural RFs. Lines at 45 and 90 degrees are not orthogonal; filters to detect edges at those orientations cannot be orthogonal. Consequently, the set of kernels can not simultaneously conserve energy and sum to the identity transform. An alternative approach would have been to use a basic set of kernels oriented at 0 and 90 degrees.

The absolute value function was used as the nonlinear transformation in the example analysis. In the multi-resolution representation at the Nyquist limit, the phase at each location is either $-\pi/2$ or $\pi/2$, so the sine of the phase is either -1 or 1. Consequently, taking the absolute value yields the local magnitude of the signal.

The orientation kernels sum to the identity transform, so the sum of the absolute value of the orientation filters applied to the band-pass image is equal to the absolute value of the sum of the orientation filters applied to the band-pass image. Squaring to use the energy envelope instead of the amplitude envelope would have been the correct nonlinear formulation if we had chosen a set of kernels that conserved energy.

II.4.1.1 Scene Synthesis Algorithm

The scene synthesis algorithm creates new content inside the target region in two passes, one for each stage of processing in the spatial vision model. Each processing step uses a spatially nonstationary

multi-resolution low-pass filtering operation to extrapolate the background characteristics (the intensity and the intensity modulation amplitude envelope) into the target region.

The first pass extrapolates the local background intensity into the target region. It creates a new image directly and is entirely deterministic. After the first pass the content in the target region converges to the local background intensity at each point on the boundary on each multi-resolution spatial frequency band.

The second pass extrapolates the amplitude envelope on each orientation and spatial frequency band into the target region. The amplitude envelope does not contain phase information, and phase information is required to create a realization, or visualization, of target region content. Near the boundary, i.e., within one-half wavelength at each multi-resolution spatial frequency band, the phase is determined by the constraint that the characteristics of the synthetic content converge to those of the local background. Deep inside the target region, i.e., more than one-half wavelength, the phase is indeterminate. Phase information is needed to create an image realization to visualize the background characteristics. In the deep interior, synthetic phase is created via a pseudo-random process. Once the phase information is incorporated, the resulting band-pass planes are recombined to create an image by inverting the orientation filtering and the multi-resolution spatial band-pass filtering as described in the preceding section. After the second pass, the new content converges to both the local background intensity and the local amplitude envelope of the background intensity modulation, at each point on the boundary on each orientation and spatial frequency band:

$$\text{Pass1Image} = \text{Extrapolation}(\text{Original Image}, \text{Mask}) \tag{17}$$

$$\text{InnerMask}_i = \{\ 1 \text{ when } L_i(\text{Mask}) > \varepsilon, \text{ else } 0\} \tag{18}$$

$$\text{NewSign}_{i\theta} = \text{InnerMask}_i * \text{Sign}(\ C_{i\theta}(\ \text{RandomImage}\ )\ ) + (\ 1 - \text{InnerMask}_i\ ) * \text{Sign}(\ C_{i\theta}(\ \text{Pass1Img}\ )\ ) \tag{19}$$

$$\text{NewAbsEnvelope}_{i\theta} = \text{Extrapolation}(\ |\ C_{i\theta}(\text{Pass1Img})\ |, \text{InnerMask}_i\ ) \tag{20}$$

$$\text{NewB}_{i\theta}(\text{Input}, \text{Mask}) = \text{NewSign}_{i\theta} * \text{NewAbsEnvelope}_{i\theta} \tag{21}$$

The extrapolation algorithm avoids use of any content from inside the target region, so that synthetic content has the characteristics of the background only. The extrapolation computes the expected value of the local background signal at each multi-resolution scale:

$$N_0(\text{Input}, \text{Mask}) = \text{Input} \times \text{Mask} \tag{22}$$

$$N_{i+1}(\text{Input}, \text{Mask}) = L_i(\text{Input} * \text{Mask}) / L_i(\text{Mask} + \varepsilon) \tag{23}$$

$$M_i(\text{Img}, \text{Mask}) = L_i(\ \text{Mask}\ ) * N_i(\text{Img}, \text{Mask}) + (1 - L_i(\text{Mask})\ ) * \text{Expand}(\ N_{i+1}(\text{Img}, \text{Mask})\ ) \tag{24}$$

$$\text{Extrapolation}(\text{Input}, \text{Mask}) = M_0(\text{Input}, \text{Mask}) \tag{25}$$

where Input is the input image, Mask is an image containing zero inside the target region and one outside the target region, and ε is a small number to prevent zero divides. For the second stage, Input is a multi-resolution amplitude envelope pyramid, and Mask is a multi-resolution mask pyramid containing zero deep inside the target region and one elsewhere.

II.4.1.2 Example Background Bias Image Results

We selected a complex image containing a variety of conditions under which to test the algorithm. The image is a still life. It contains a variety of objects with different shapes, textures, and

luminance. It contains complex shadowing, layering, and juxtaposition. Some objects have regular shapes, some are irregular. Some are large, and some are small. The overall organization of the scene is irregular, but recognizable as a classic still-life form. The still-life image is shown in figure 9.

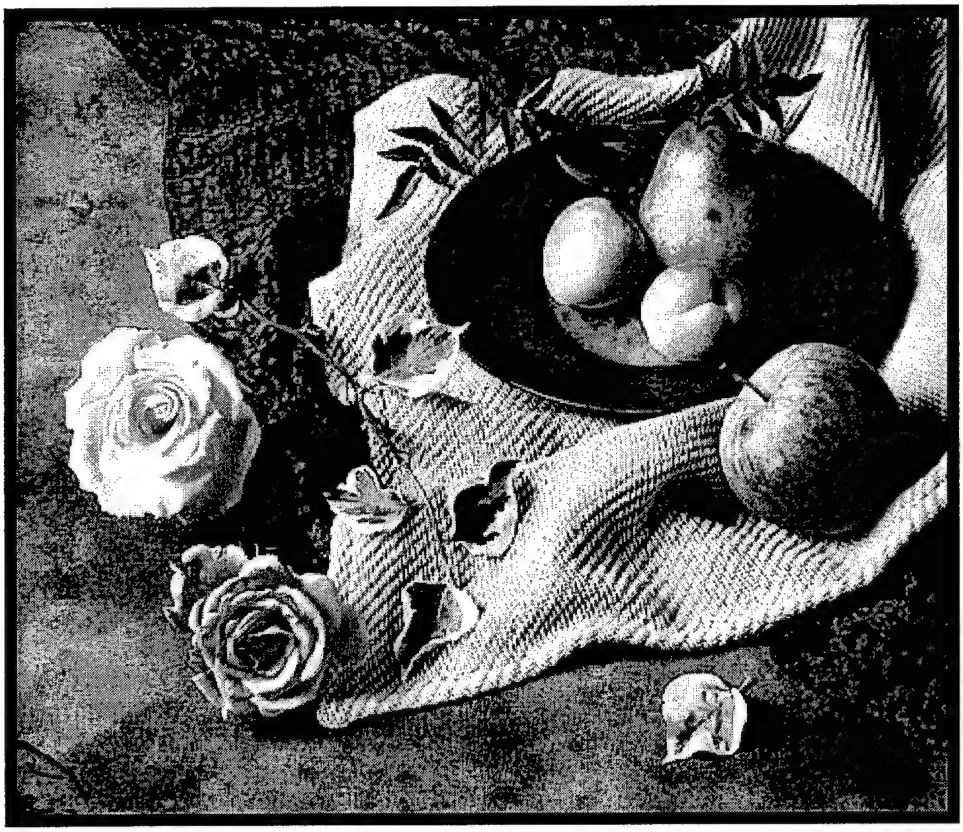

Fig. 9. Original image

We designated a dozen target regions in the image. The target regions were selected with a variety of shape characteristics—some were rectangular, some were circular, some highly irregular. Some target regions were narrow and some were wide. The positions of the target regions were selected to sample a variety of simple and complex local background conditions. Some of the target regions are surrounded by relatively uniform luminance and texture, while others border on different luminance and textures. Some of the target regions appear against a single background object, while others straddle several background objects. In some cases the obscured background object borders are regular and easily anticipated, and in other cases are irregular and less easily anticipated. The target region mask is shown in figure 10.

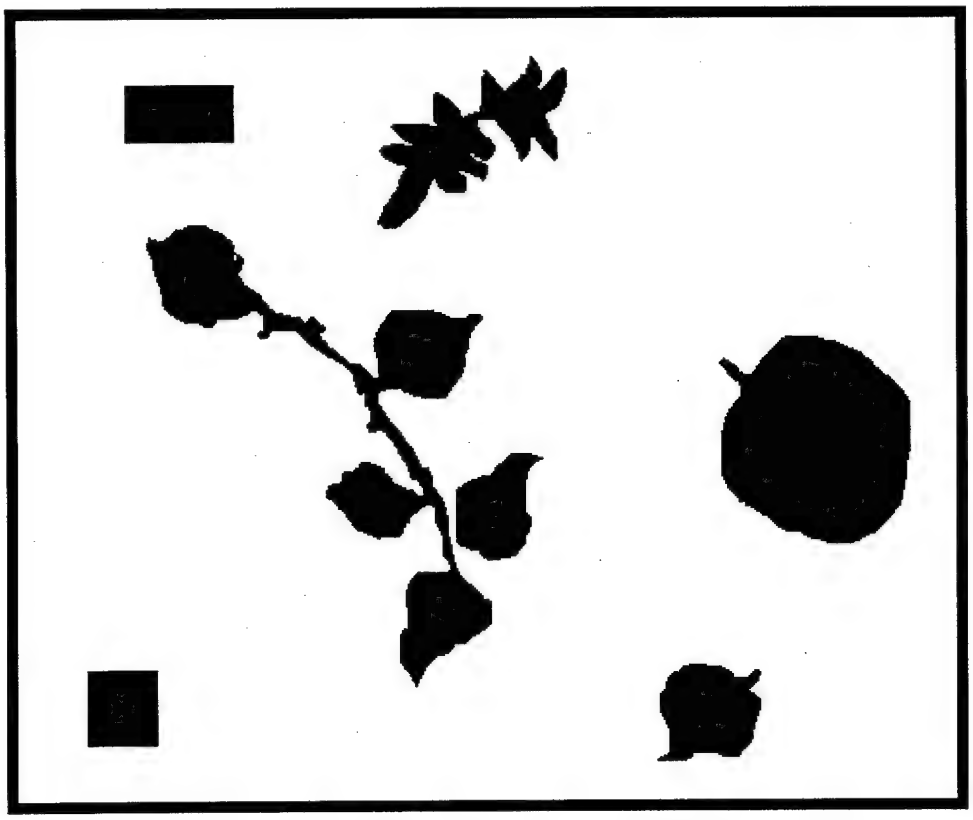

Fig. 10. Mask image

Figure 11 shows the results of the first pass of the algorithm. This is the background bias image for computing the first-stage (luminance gradient) metric component. The target region is matched to the local luminance on each spatial frequency band. Luminance gradients are extrapolated into the target regions.

The target regions are easily distinguished by their lack of internal texture and by very local contrasts against background areas of high modulation. Except where the local background is relatively uniform, the region boundaries are visible due to texture gradients. For the most part, luminance gradients do not contribute to the boundary definition. One exception is the upper right arc of the large apple. In this instance the target mask missed part of the apple, and it is the missing sliver of apple that creates the boundary appearance.

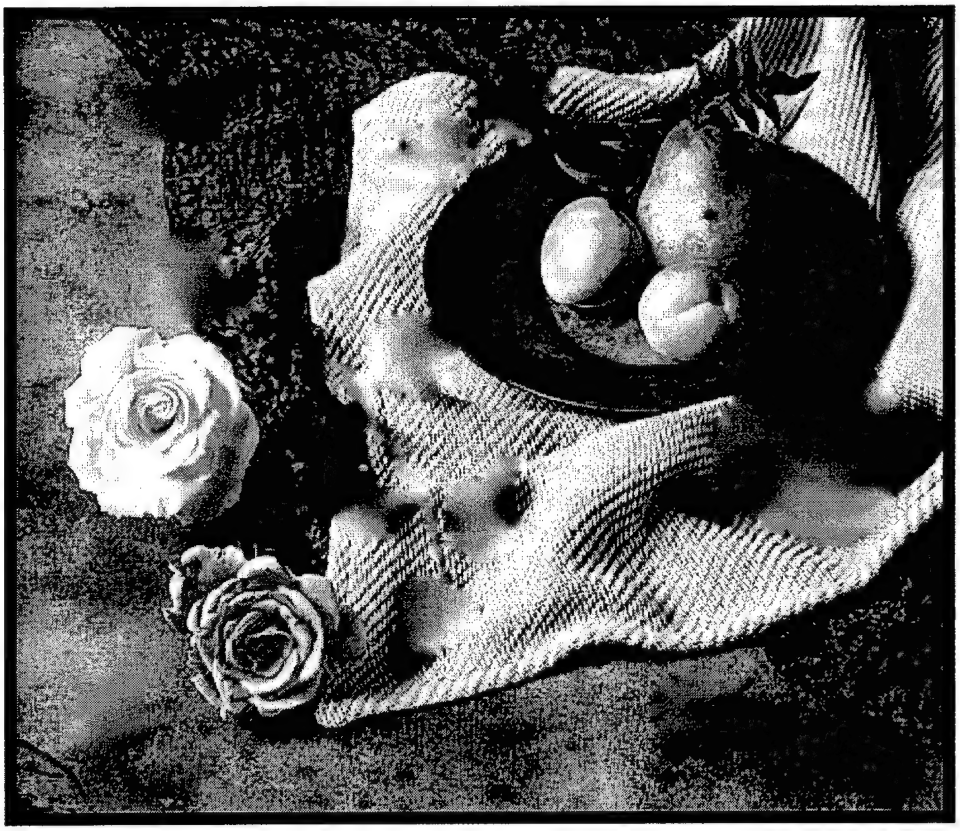

Fig. 11. Luminance gradient background bias image

Figure 12 shows the results of the second pass of the algorithm. This is the background bias image for computing the second-stage (texture gradient) metric component. The target region matches the local intensity and amplitude envelope at each point on the boundary. This pass significantly reduces the visibility of the boundary for most of the target regions. With the exception of the large apple, it is very difficult to determine the shapes of the target regions. Portions of the boundary of the apple are still distinct. The apple is a problem because it is large, its boundary is regular, and it borders on a variety of different objects.

The pseudo-random texture interior is less distinctive than the uniform interiors of the first pass. Many of the target regions are difficult to distinguish from the background, especially when the surrounding texture is relatively homogeneous. On the interior of the large target regions, the amplitude envelope is relatively uniform, and this uniformity of texture is noticeable. In some cases, boundaries between background objects are extrapolated into the target regions, but in other cases the boundaries are simply blurred.

The boundary of the plate is blurred where tip of leaf crosses it in the center of the image. The boundary of upper left leaf (just above the rose) is indistinct, but the replacement content is distinctly not representative of the scene. The gradient between the table cloth and the table top is clearly not extrapolated through region. This is due in part to the presence of the bright flower just below the leaf. In
essence, the algorithm blends the characteristics of the tablecloth, the tabletop, and the boundaries between them. It creates a mixture of the local characteristics, which ends up looking like none of the particular objects or boundaries.

The algorithm does not "know" about objects, boundaries, and textures; it treats the image as a stochastic process with statistical properties. The algorithm does not "know" to extrapolate only from

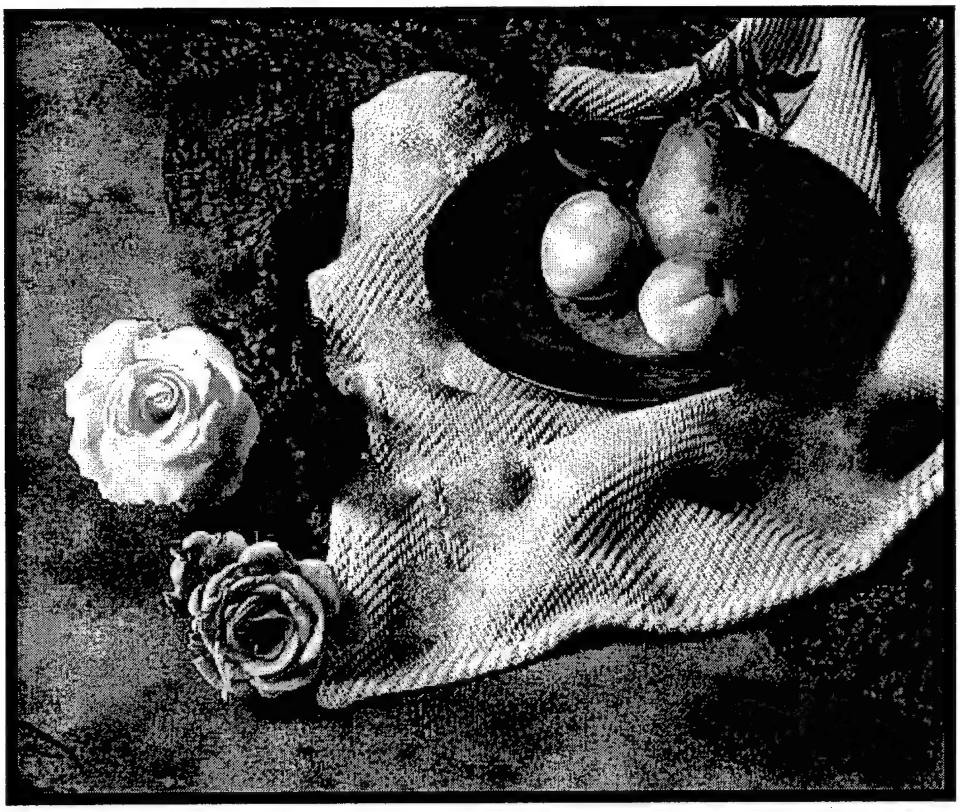

Fig. 12. Texture gradient background bias image

immediately adjacent objects, and to disregard remote objects. It does not distinguish between modulation at the boundaries between objects and modulation due to interior texture.

The appearance of texture is, in fact, due to fine boundaries between small regions within an object. Visual perception of modulation as texture rather than boundaries is influenced by prior knowledge of the objects in the scene, and their relative size and scale. The distinction is not necessarily an inherent property of the image, but depends on what we are looking for in the image. When we are parsing the scene into large objects, the tablecloth has a texture and the apple has a texture. If we were looking for small details, we would parse the scene into many small regions. The spots on the apple skin which create the texture can also be viewed as individual objects, and the weave of the fabric can also be seen as individual fibers.

II.4.1.3 Findings and Recommendations

The algorithm was effective at reducing the visibility boundary and region shape in most cases. This indicates that the background characterization provides a good reference to use in computing the perceptibility of the external shape of a target region. The algorithm was, understandably, unable to affect the boundary when a portion of the target was left outside the target region.

The algorithm did not fully obscure the boundary when the target region simultaneously met three conditions: (1) it had a simple, smooth, regular shape; (2) it was adjacent to different background components with large contrasts and texture gradients, and sharp boundaries between them; and (3) it was large with respect to the adjacent background features.

These observations suggest that the predictive model of shape perception will need to account for the complexity of the shape of the target region, e.g., to employ a shape complexity metric, and will need to account for the combination of size and shape of the region.

The algorithm always treats modulation as a statistical, stochastic process, and does not recognize that sometimes there is a deterministic, discrete component associated with object boundaries. The algorithm does not distinguish between modulation which should be treated as a boundary between objects, and modulation which should be treated as the texture of an object. It is possible that scene segmentation can be used in the local background characterization process to improve sensitivity to the discrete components.

The algorithm was moderately effective at creating synthetic target content which was not visibly distinct from the background in most of the cases. This indicates that in many cases the background characterization is also a useful reference for computing how distinctive a target is from its surroundings. It was somewhat less effective at suppressing the distinctiveness of the target regions than it was at suppressing the shapes of their boundaries.

The most notable exception is when the target occludes a boundary between two background regions, and at the same time is near other large contrasts and gradients. The algorithm does not "know" that it should only respond to one or another of the features, and not to average their characteristics. The problem is that when the distribution of surrounding background characteristics is bimodal the extrapolation into the target may be between the modes of the distribution and not resemble either. Once again, it is possible that scene segmentation can be used in the local background characterization process to improve performance in this case.

Filling in large target regions so that they are not distinguishable from their surroundings requires more than filling with statistical patterns and textures. It requires generating the appearance of decoys which make the target look like another object or objects. This requires model-based or knowledge-based generation of scene content, not merely statistical pattern analysis.

II.4.2 Information Metric Model: Boundary Information from Intensity and Texture Gradients

This section describes and illustrates the core elements of the information metric model. These algorithms compute the boundary information from first-stage (intensity) and second-stage (texture) gradients. These algorithms are integrated with the background bias image generation methods described in section II.4.1. These results are very significant because they tangibly demonstrate how the information metric model performs for high- and low-intensity gradient and contrast gradient regions in complex heterogeneous scenes.

Figures 13, 14 and 15 illustrate the first-stage (luminance) modulation amplitude envelope for the first-stage and second-stage background bias images, and the highest and second-highest spatial frequencies. We see in figures 13 and 15-left that the magnitude of the modulation inside the perimeter of the target regions is very low, except where gradients between portions of the local background bleed into the target. These images illustrate the analytical explanation for the observation that target-to-background contrast contributes little to boundary definition in the first-stage background bias image. The figures also show the high amplitude around the boundaries of the other objects in the scene, e.g., the peaches on the plate, flowers, etc. Figures 14 and 15-right show the modulation amplitude on the interior of the target regions for the second-stage background bias image. This demonstrates that the pseudo-random texture inside the target regions will produce a small but nonzero first stage measure of integration around the target region boundaries. These illustrations indicate that integrating the first-stage amplitude envelope around the target boundary will measure the contribution of target-background contrast to boundary perception. They also show that the integral around the interior of the target regions

in the first-stage amplitude envelope of the first-stage background bias image yields the minimum possible value, given the surrounding image.

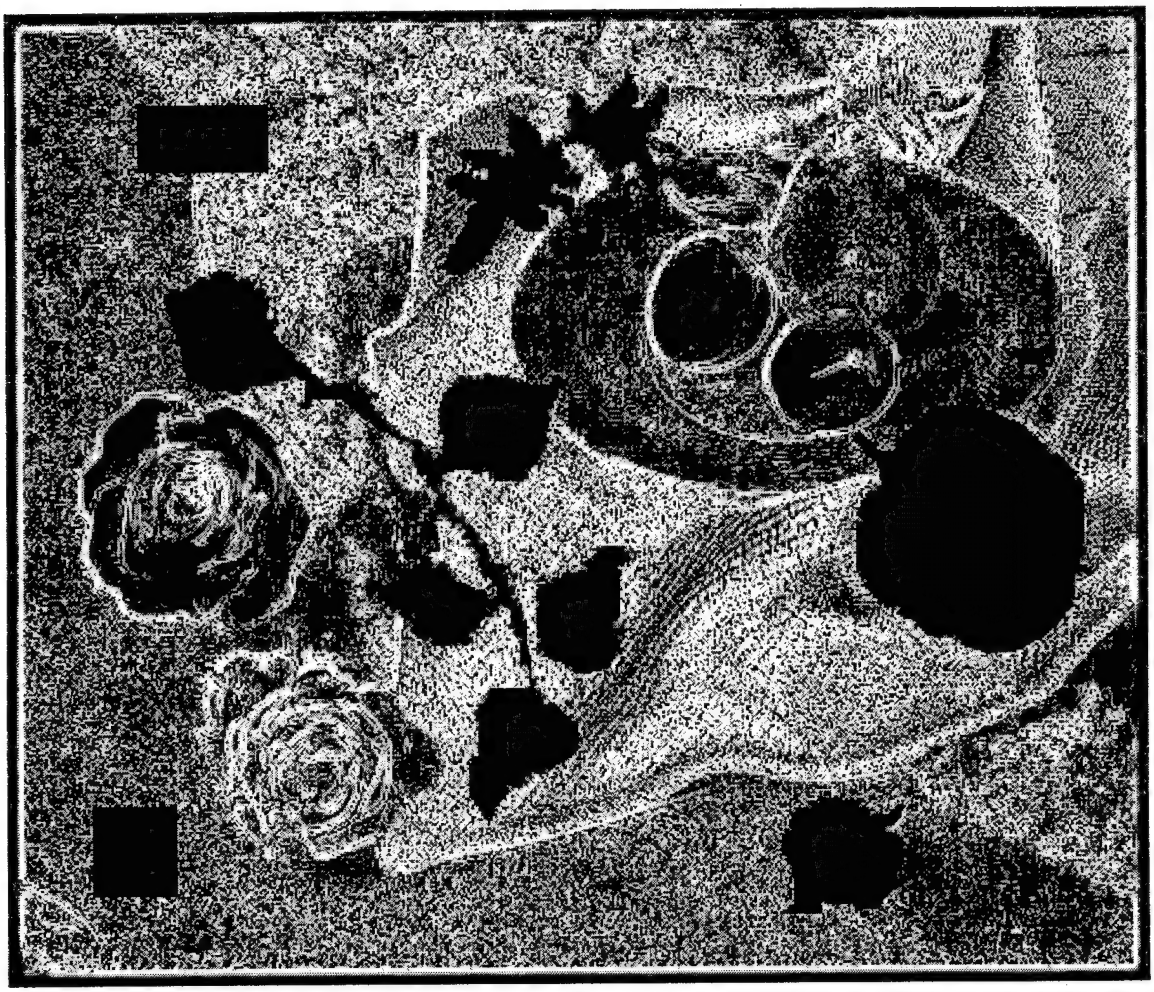

Fig. 13. First-stage (luminance modulation) amplitude envelope on the highest spatial frequency band of the first-stage (luminance modulation) background bias image

Fig. 14. First-stage (luminance modulation) amplitude envelope on the highest spatial frequency band of the second-stage (texture) background bias image

Fig. 15. First-stage (luminance modulation) amplitude envelope on the second-highest spatial frequency band for the first-stage (luminance modulation) (left) and the second-stage (texture) (right) background bias images

Figures 16, 17 and 18 illustrate the corresponding results for the second-stage (texture) modulation amplitude envelope. In figures 16 and 18-left, we see the characteristic bright ring on the inside of target regions, showing the magnitude of the texture gradient and its contribution to the boundary information metric. These images show the analytical explanation for the observation that the texture gradients reveal some of the shape of the target regions in the first-stage background bias image. The figures also show the high amplitude around the boundaries of the other objects with texture gradients, e.g., the peaches on the plate, flowers, etc. Figures 17 and 18-right show the texture modulation amplitude on the interior of the target regions for the second-stage background bias image. These illustrations show that integrating the second-stage amplitude envelope around the target boundary will measure the contribution of target-background texture contrast to boundary perception. They also show that the integral around the interior of the target regions in the second-stage amplitude envelope of the second-stage background bias image yields the minimum possible value, given the surrounding image.

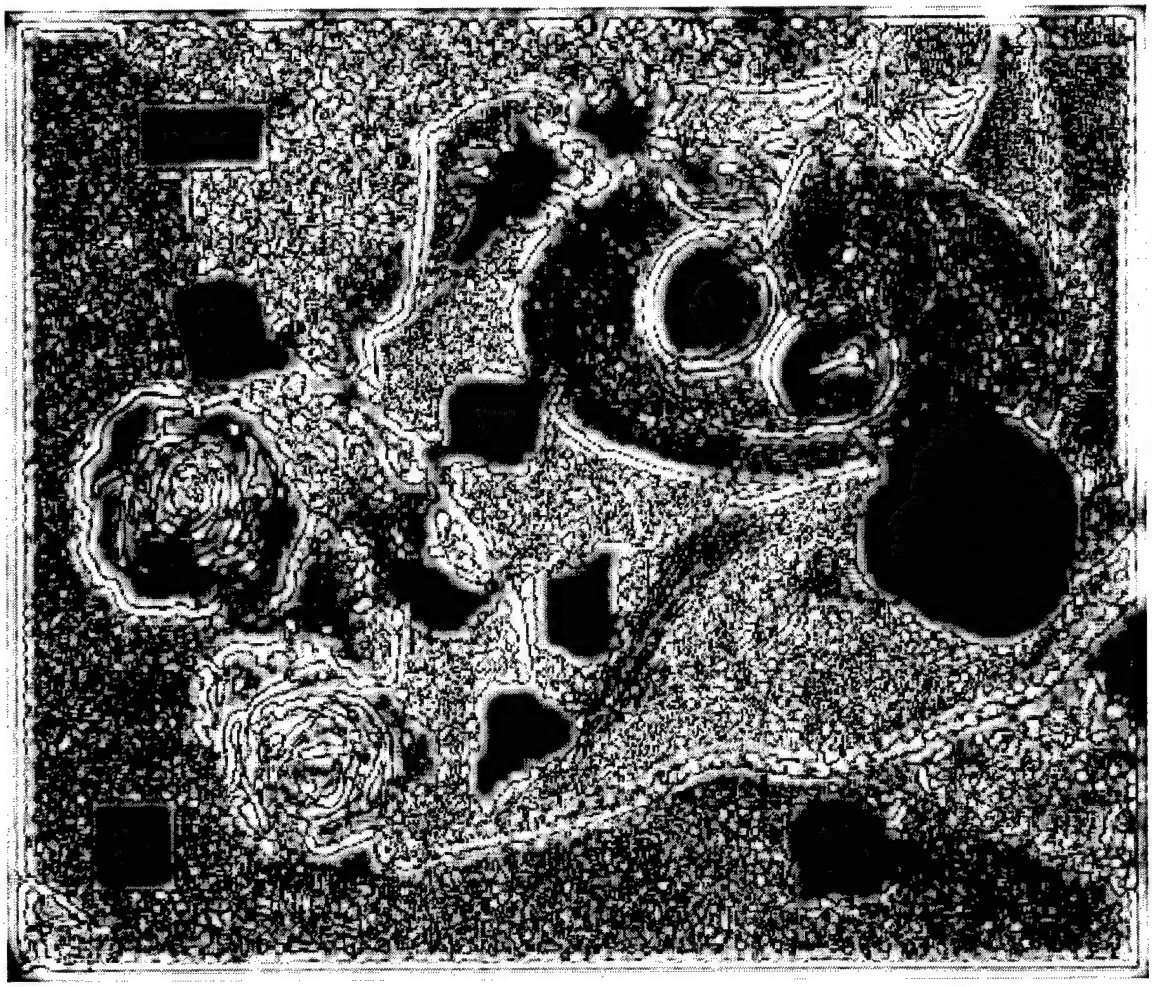

Fig. 16. Second-stage (texture) amplitude envelope on the highest spatial frequency band of the first-stage (luminance modulation) background bias image

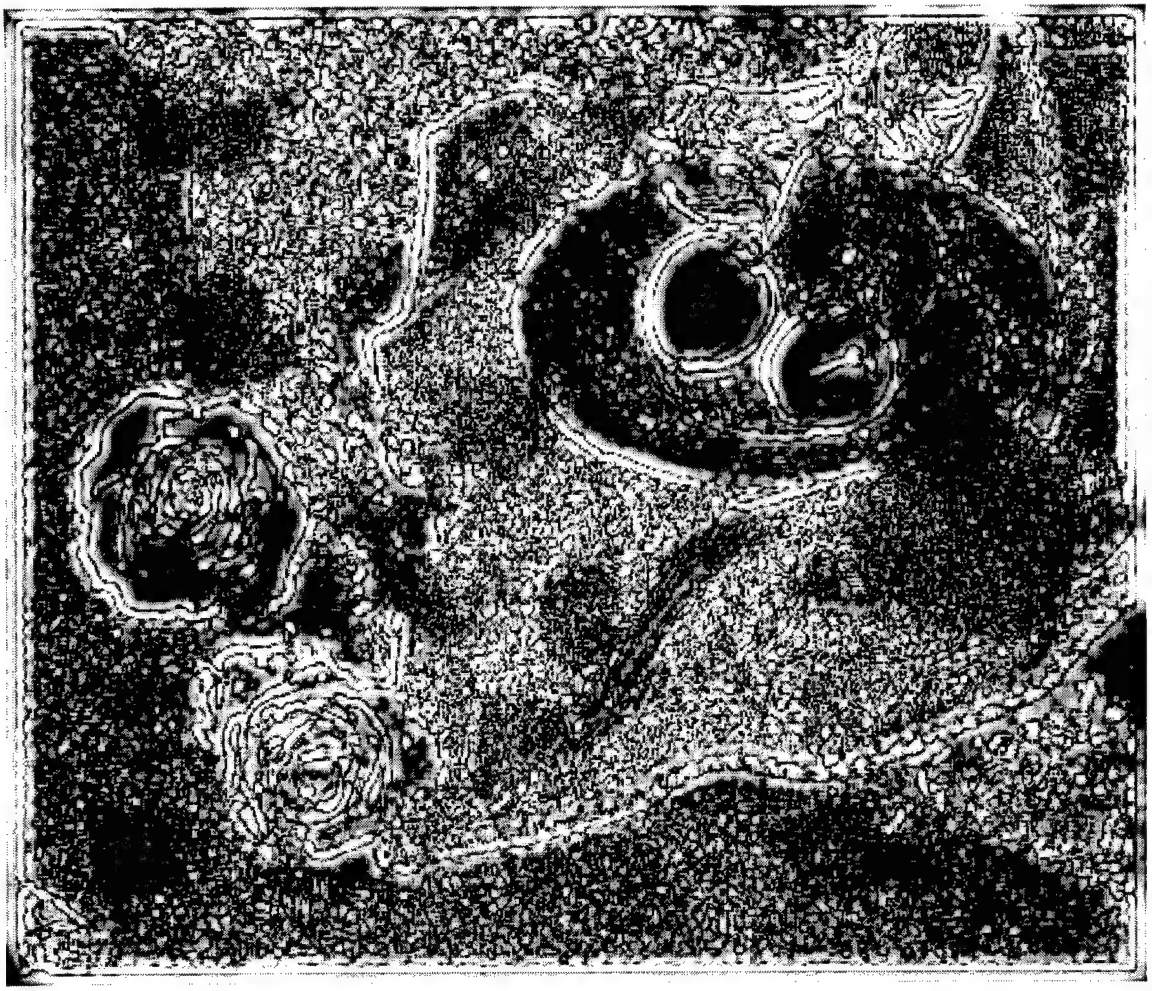

Fig. 17. Second-stage (texture) amplitude envelope on the highest spatial frequency band of the second-stage (texture) background bias image

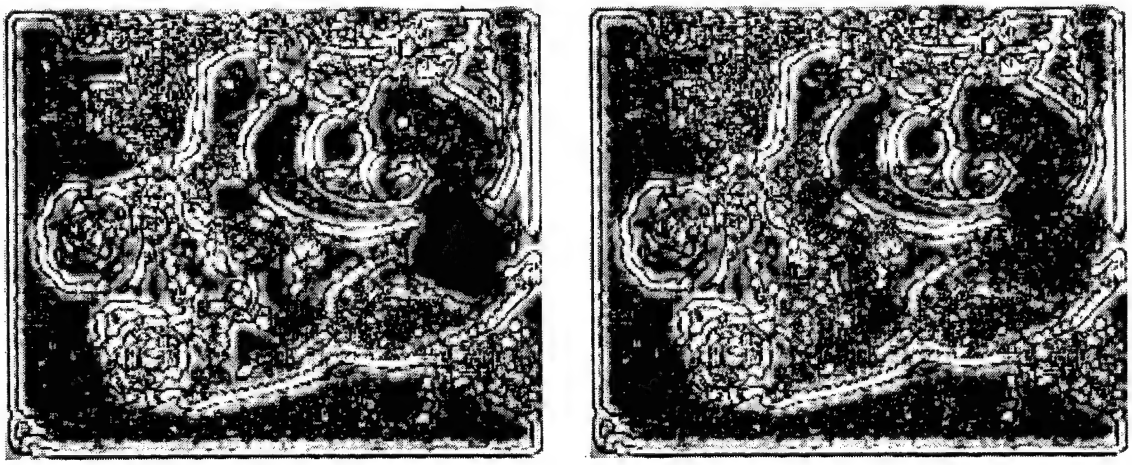

Fig. 18. Second-stage (texture) amplitude envelope on the highest spatial frequency band of the first-stage (luminance) (left) and second-stage (texture) (right) background bias images

II.4.3 Image Segmentation

This section illustrates multi-resolution image segmentation algorithms. The initial concept was to segment the image at the zero-crossings on each multi-resolution band-pass image plane. For purposes of the illustration, we focused on the first-stage band-pass of the intensity image, rather than the second-stage amplitude envelope images. Segmenting on each multi-resolution plane captures different features at the different resolution levels. Segmenting at the zero-crossings separates regions which, at each multi-resolution plane, are lighter than their local surround from those which are darker than their local surround. Marr [1982] posited a multi-scale representation in his theory of edge detection in which the zero-crossings define the region boundaries at each level of spatial resolution. The zero-crossings are the boundaries between the positive and negative regions of each multi-resolution band-pass plane.

Before elaborating on this approach, it is worth noting that image segmentation is also a topic in the digital image processing literature. Gonzalez and Wintz [1995: 331-90] review these techniques. The techniques did not offer much promise for complex heterogeneous scenes such as the sample image. Although mathematically interesting, the examples were restricted to simple targets and backgrounds with simple noise processes.

The value of the information contained in just the zero-crossings is illustrated in figure 19. This

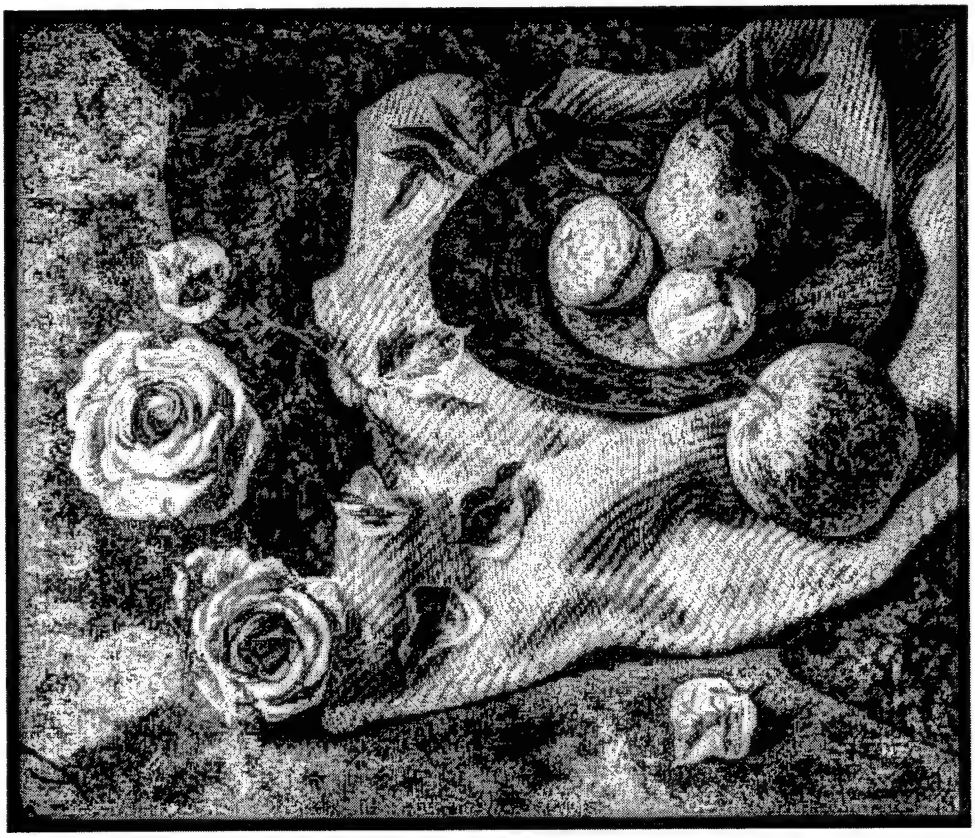

Fig. 19. Multi-resolution magnitude equalization

figure was constructed by segmenting each plane of the multi-resolution image into its positive and negative components. Pixel values of +1 were assigned to pixels where the sign of the band-passed image was positive, and -1 where it was negative. These image planes contained only the multi-resolution phase information and no magnitude information. The image planes were then expanded and

summed to construct the image in figure 19 using the standard VPM inverse band-pass operation to represent the phase content of all the band-pass channels in a single image. The algorithm is

$$X_{fmax} = \text{Sign}(B_{fmax}) \quad (26)$$

$$X_f = \text{Sign}(B_f) + \text{Expand}(X_{f+1}) \quad (27)$$

$$\text{Multi-resolution magnitude equalization image} = X_0 \quad (28)$$

where B_f denotes the multi-resolution band-pass image planes.

The result is an enhancement of local detail at all spatial scales. In figure 19 we are able to see the heretofore invisible pattern on the plate, and mottled texture of the fruit without losing any of the larger features of the shadows, and relative darkness of the tablecloth. Large, medium, and small features are all equally visible. Large and small luminance gradients are also equally visible. It is worth noting that the band-pass images could have been passed through a "dead-band" filter to eliminate small modulations relative to the eye noise magnitude before segmentation.

This example illustrates the potential value of the information in the multi-resolution sign images. Unfortunately, the multi-resolution zero-crossings do not turn out to be directly useful in complex heterogeneous images which are the focus of this research. They tend to divide complex objects, not to segregate whole objects, and they fail to distinguish texture from boundaries. There are many boundaries between light and dark regions in heterogeneous scenes which do not correspond to anything that we want to call objects. These boundaries create zero-crossings at multiple levels of resolution, just like "true" boundaries.

In heterogeneous scenes, the region segments often do not correspond to the objects in the scene. Most of the objects in the sample scene contain both light and dark regions. The zero-crossings divide the objects into different regions, rather than isolate unified objects. In some cases, e.g., the flowers and leaves, the internal zero-crossings delineate boundaries between different regions of the same object. At scales smaller than the objects of interest, the zero-crossings tend to break up the objects rather than consolidate them. This is because the objects in the scene (e.g., the rose flowers, or the leaves, or the plate) consist of both light and dark regions. At scales larger than the objects of interest, the zero-crossings tend to consolidate different objects, and smooth the boundaries thus losing important shape detail. These problems are exacerbated if we do not know the scale of the objects of interest in advance, or if the objects of interest are at difference scales in the image (e.g., targets at different ranges).

In other cases, e.g., the woven cloth and the mottled fruit, the zero-crossings correspond to the internal texture of the objects. When examined closely, the apparent texture actually is created by many small regions, with boundaries between them. A grassy field has the texture of grass, but also consists of many individual blades of grass. The mottled texture is created by individual blotches. The woven texture is created by many small shadows. The zero-crossings cannot directly distinguish between modulation that we perceive as a region of texture, and modulation that we perceive as boundaries between light and dark regions.

II.4.3.1 Identifying Object Boundaries

We attempted to improve the object segmentation performance by selecting the scale for the initial segmentation, disregarding all lower frequency information, then using the higher frequency information to refine the location of the boundaries defined at the basic spatial frequency. The basic idea is to select the scale appropriate to the size of the objects of interest for the initial segmentation, then to sharpen the boundary definition by using higher spatial frequency information. The equations for the algorithm are:

$$X_{fmax} = \text{Sign}(B_{fmax}) \quad (29)$$

$$X_f = Z_f * \text{Sign}(B_f) + (1 - Z_f) * \text{Sign}(\text{Expand}(X_{f+1})) \quad (30)$$

$$\text{If } (B_f = 0, \text{ or Expand}(X_{f+1}) > \varepsilon), Z_f = 0 \text{ else } Z_f = 1 \quad (31)$$

$$\text{Segregation Image} = X_{fmin} \quad (32)$$

B_f denotes the multi-resolution band-pass image planes. The free parameters are fmax, the initial (lowest) spatial frequency, fmin, the final (highest) spatial frequency, and ε , the decision threshold specifying when to sharpen with the sign of the higher frequency band-pass data versus using the expanded lower frequency intermediate results.

Figure 20 shows the fully expanded segregation using only the fourth level of the Laplacian pyramid. This starting point was chosen because the scale corresponds approximately to the size of the plate, flowers and pieces of fruit that we would like to call the objects in the scene. However many of the objects of interest are connected and not individually segmented.

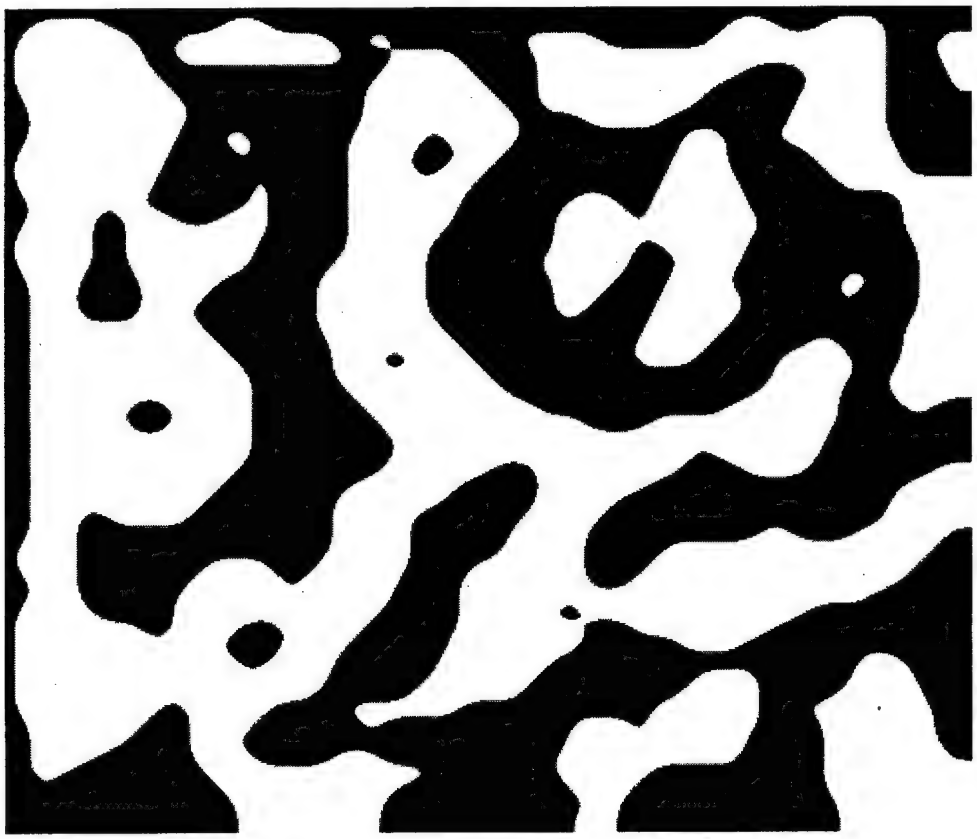

Fig. 20. Initial segmentation image on multi-resolution plane 3

Figures 21 , 22, and 23 show the sequence of sharpened segregation images with large value of the decision threshold ($\varepsilon = 0.99$). The image sequence shows improved segmentation of the objects in the scene, but at the same time there is greater segregation of features in the image which we do not necessarily want to call objects, e.g., the details of the woven cloth texture. There is no spatial frequency cutoff that achieves one without the other.

Fig. 21. First refinement ($\varepsilon = 0.99$)

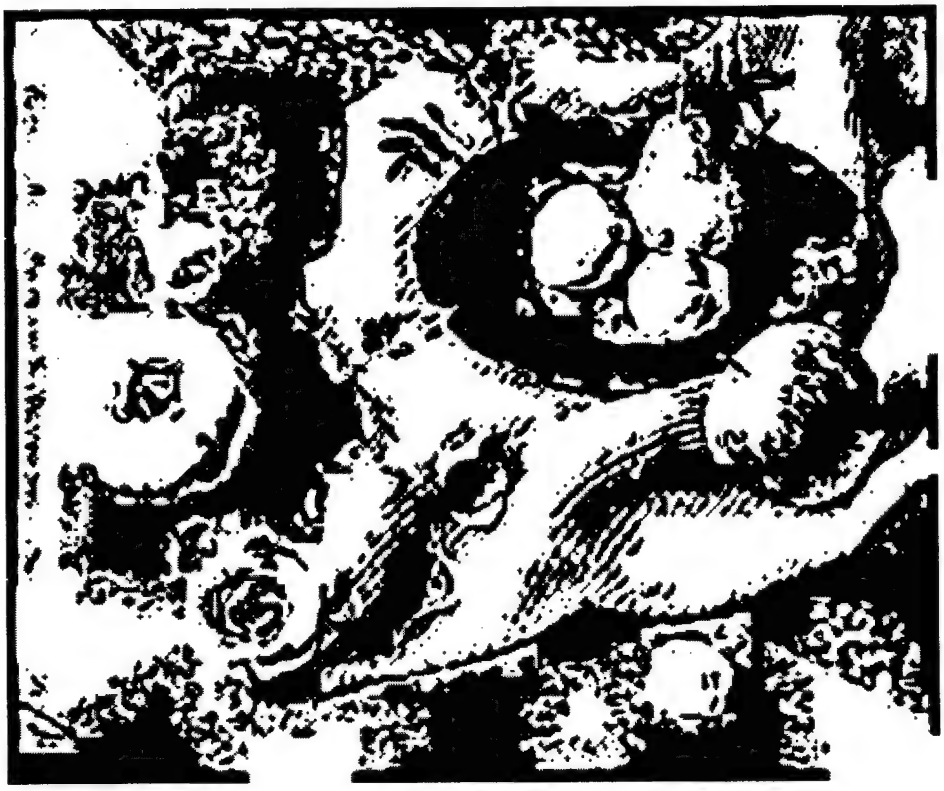

Fig. 22. Second refinement ($\varepsilon = 0.99$)

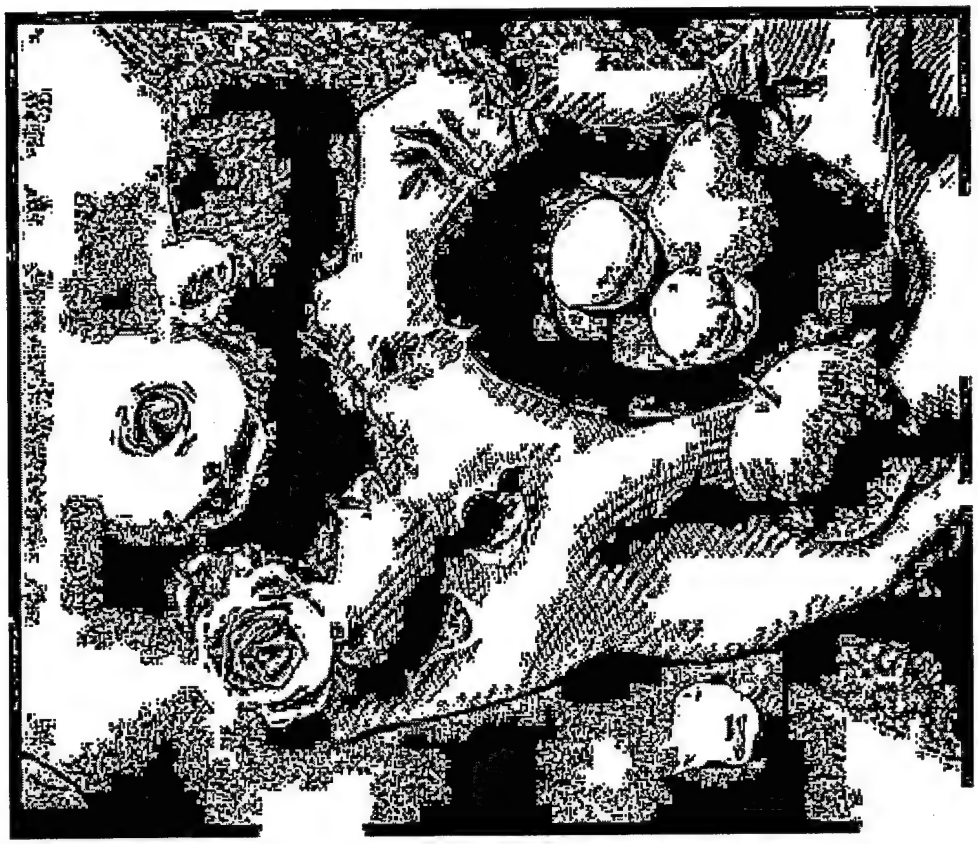

Fig. 23. Third and final refinement ($\varepsilon = 0.99$)

Figures 24, 25, and 26 show the sequence of sharpened segregation images with small value of the decision threshold ($\varepsilon = 0.5$). The sharpened boundaries stay much closer to the original, with little extraneous segmentation, but at the same time do little to improve the segmentation of the objects of interest.

The pieces of fruit in the center of the plate are joined into one region in the initial segmentation (figure 20), but are separated in the subsequent sharpened segmentations. They are segmented much better with epsilon equal to 0.99 (which introduces spurious texture) than with epsilon equal to 0.50 (which better rejects spurious texture).

The results are interesting, but do not demonstrate a reliable method for object segmentation. If we set the width threshold high enough to separate distinct objects which were consolidated at the lower level of resolution, we see the "fractile" nature of the zero-crossings emerge. They are not smooth and small features emerge. If we set the threshold low to reject these features, then we are unable to separate objects that we want to separate.

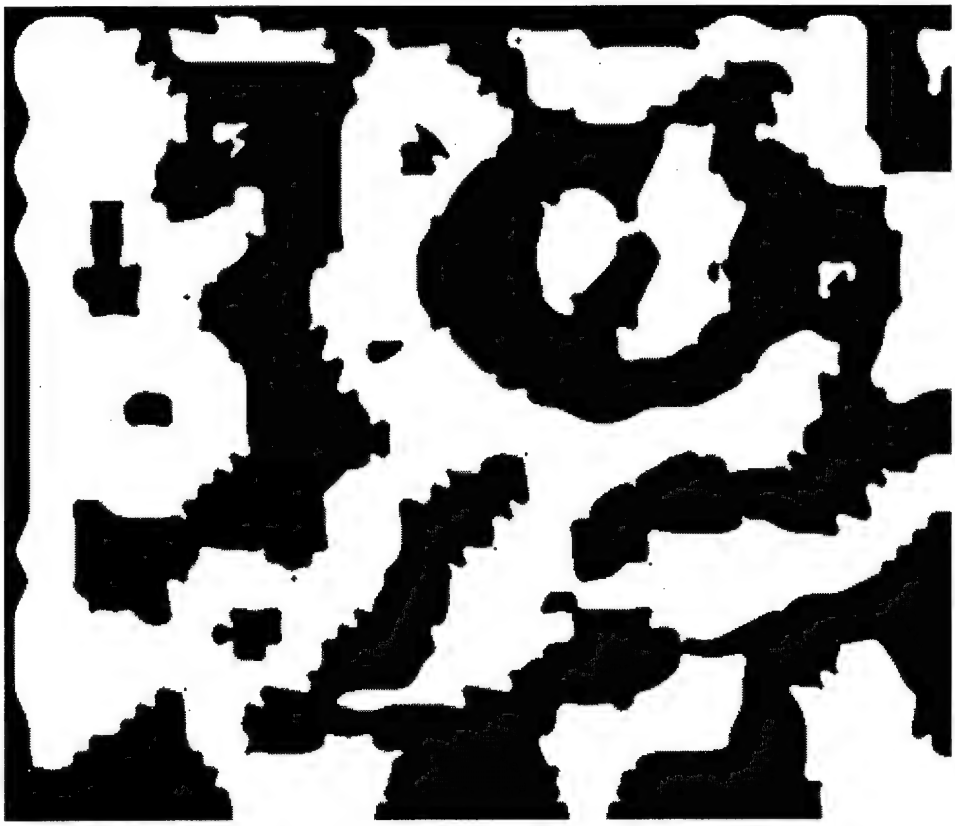

Fig. 24. First refinement ($\varepsilon = 0.50$)

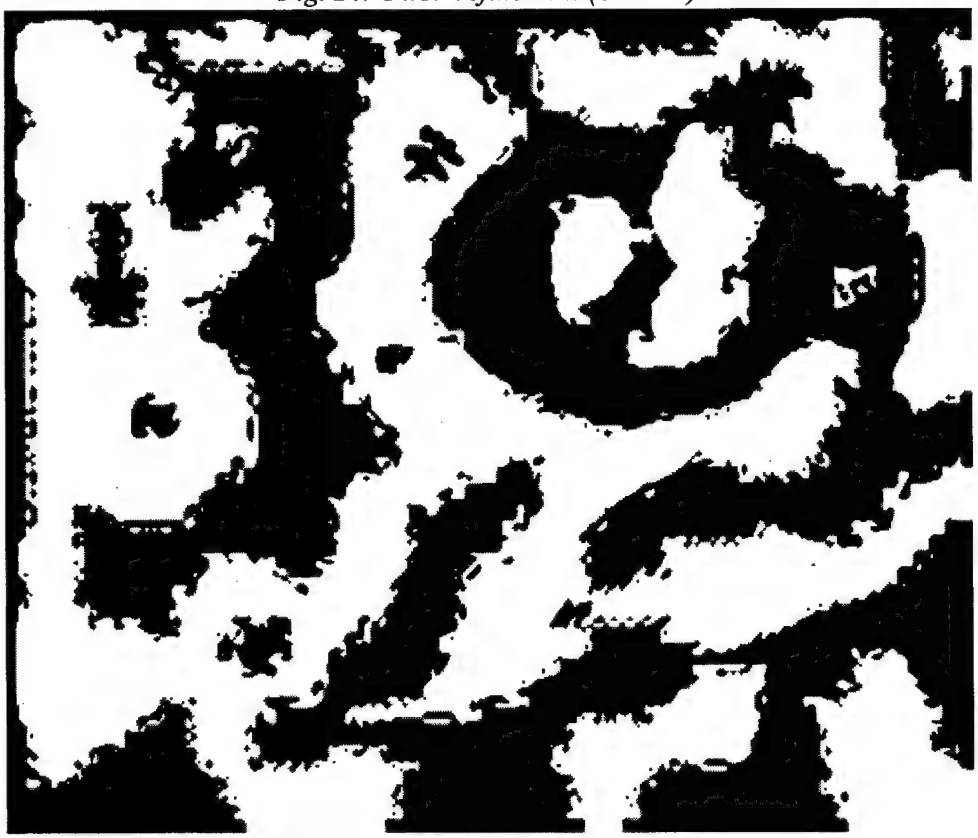

Fig. 25. Second refinement ($\varepsilon = 0.50$)

Fig. 26. Third and final refinement (ε =0.50)

II.4.3.2 Distinguishing Object Boundaries from Internal Texture

The preceding results again highlight the difficulty of distinguishing boundaries from texture, even using multi-resolution correlation as the basis to define boundaries. The problem is that the appearance of texture is created by a pattern of boundaries of small regions filling a larger area. Filtering to find boundaries will inevitably also select the micro-features that create the texture, unless higher level knowledge can be applied. However, they provide some valuable insight into the problem of distinguishing object boundaries from internal texture: we see that boundary detection will turn up texture, and that we want to call things texture when they seem to fill an area rather than delineate a region. Boundaries are a linear property. Texture is an area property.

Gonzalez and Wintz [1987: 414-23] review digital image processing approaches to texture. They note that there is no formal definition of what is meant by texture, or standard computational means to measure it. They present a variety of statistical, structural, and spectral approaches to quantify texture. The spectral methods of Fourier analysis (power spectral density) are very similar to the amplitude envelope on the spatial frequency band-pass channels. However, the analysis methods are all global: they assume that the images are of a homogeneous texture. The methods are inapplicable to heterogeneous images, and even if extended would require external specification of what the texture regions are. These methods did not appear to have any potential value for distinguishing modulation perceived as texture from modulation perceived as boundary.

The proposed approach to distinguishing modulation perceived as texture from modulation perceived as a boundary hinges on the fact that perceived texture fills a region, whereas perceived boundary is linear. Edges, or correlation between multi-resolution planes, create the underlying signal

which can be perceived as either texture or boundary. The inspiration for the analysis method is the algorithm to draw a line which fills space. The technique is to define a sequence of lines. The i^{th} line is drawn such that the maximum distance between it and the previous line is $1/2^i$. The result is that for any given point (x,y) and any given distance d, there is a sequence number j such that all lines in the sequence are within distance d from (x,y). This sequence of lines is dense. It fills area. Lines, or sequences of lines, which do not have this property can enclose area, but do not fill area.

This is the basic idea that we are attempting to exploit to distinguish perceived boundary from perceived texture at any given level of resolution. Where the zero-crossings are dense, we call them texture. Where the zero-crossings are not dense, we call them boundary. Of course, the size of the objects of interest determines whether a region is regarded as containing a texture or many boundaries. A grassy field can be perceived as an area with the texture of grass, or a large number of individual stalks of grass.

The basic algorithm is presented in figure 27. The first step is standard multi-resolution band-pass analysis. The second step is to segment each image plane into its positive and negative regions. The zero-crossings are the boundaries between these regions. The third step is to segment neighborhoods near the zero crossings by applying a spatial low-pass filter, taking the absolute value, and thresholding. The low pass places values of +1 and -1 away from the zero crossings, and values strictly between +1 and -1 near the zero-crossings. The absolute value causes pixels away from the zero-crossings to have value of one, and pixels near the zero crossings to have values strictly less than one. The threshold maps pixels with value above the threshold to 1 (denoting pixels away from the edge), and pixels with value below the threshold to 0 (denoting pixels near the edge). The fourth step collects close neighborhoods and rejects narrow isolated neighborhoods. It uses a second low-pass filter, to blend close regions together, then a second threshold to unambiguously segment areas near zero-crossings from those which are not. The value of the threshold needs to be high enough to reject regions which are near a single zero-crossing (and hence perceived as boundary rather than texture).

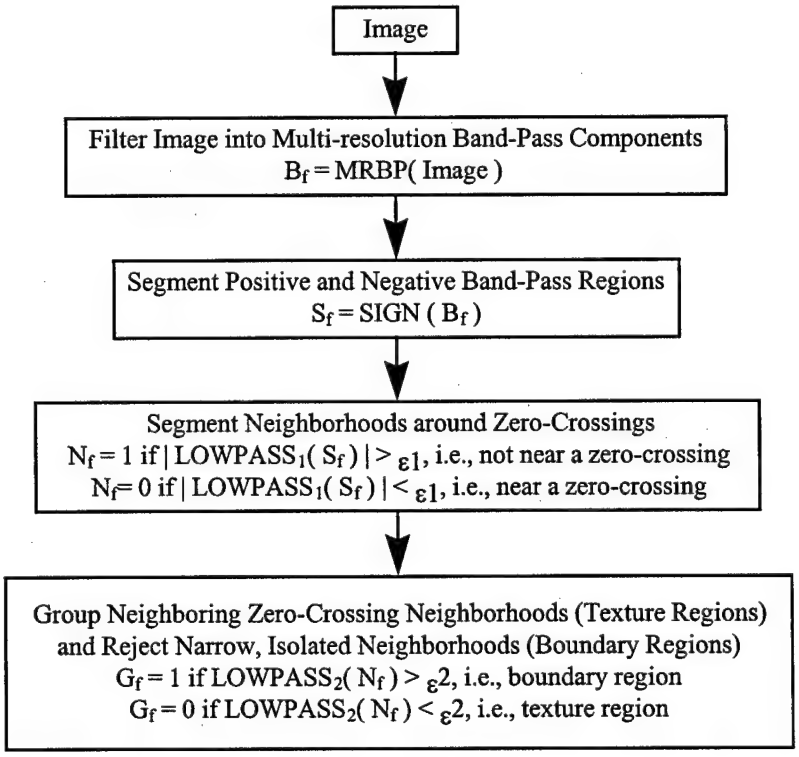

Fig. 27. Texture / Boundary Segmentation Algorithm

This provides the input to determine which regions are perceived as being filled with a texture. Compact regions of zero-crossings, i.e., in which the ratio of the area to the perimeter squared is large, will tend to be perceived as filled with a texture. Regions containing a single zero-crossing line which are perceived as a boundary will not be compact. Refinements of this algorithm and its use in image segmentation will be considered in Phase II.

Initial testing indicates that this is a very promising approach. The algorithm has four free parameters (two thresholds, and the widths of the two low-pass filters). The following illustration was created with "best initial guess" parameter values without any iteration or optimization. The potential of the algorithm is shown in figures 28 and 29. We used the algorithm to try to separate the sample image

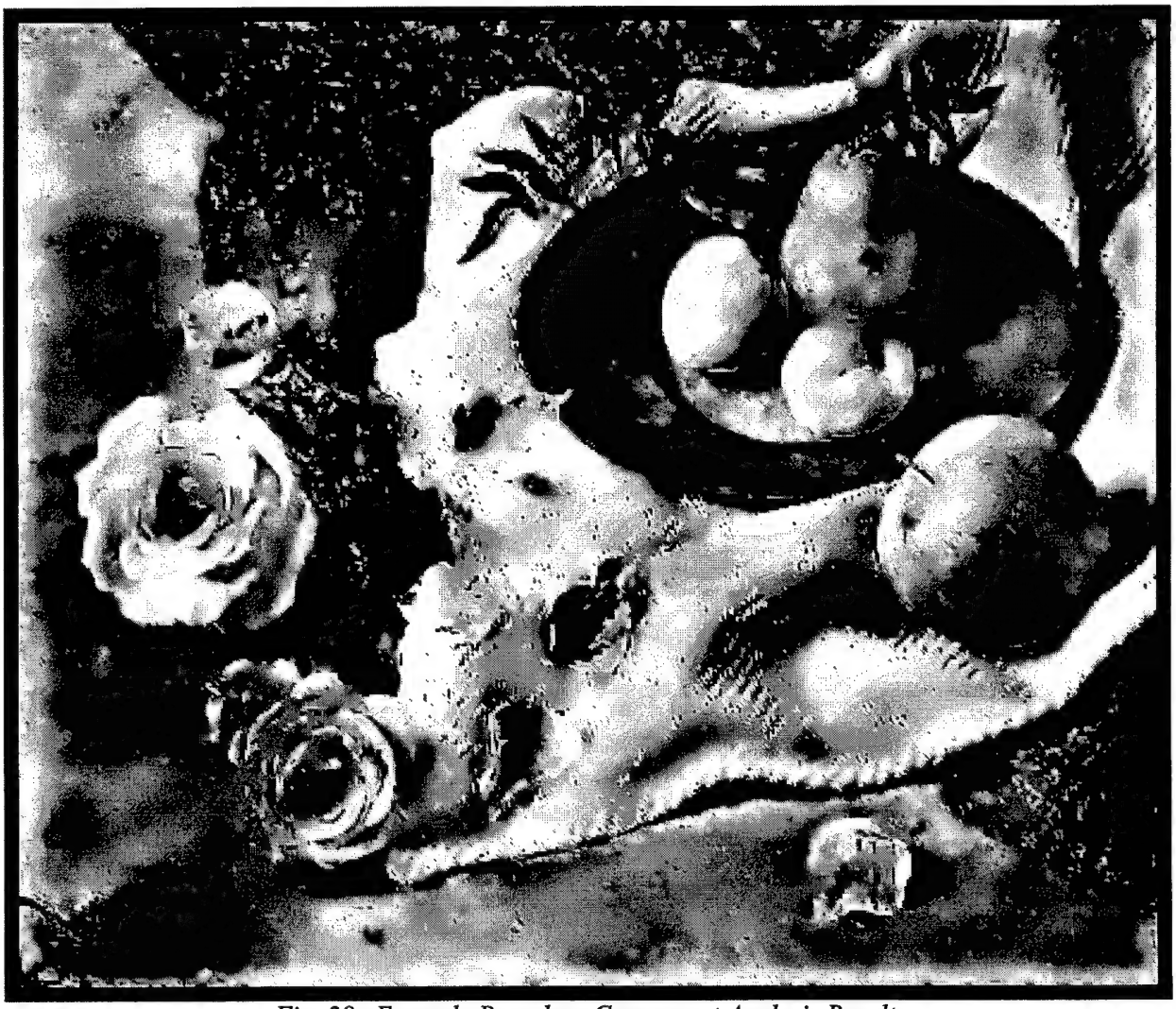

Fig. 28. Example Boundary Component Analysis Results

into its boundary and texture components. Figure 28 is the roll-up, over all spatial frequency bands, of the boundary components, and figure 29 is the roll-up of the texture components. For the most part, figure 28 shows the objects without internal texture. For the most part, the internal textures are contained in figure 29. The performance of the first cut of the algorithm is not perfect. Object

boundaries that are adjacent to a texture, e.g., the edge of the plate next to the woven cloth, are picked up on the

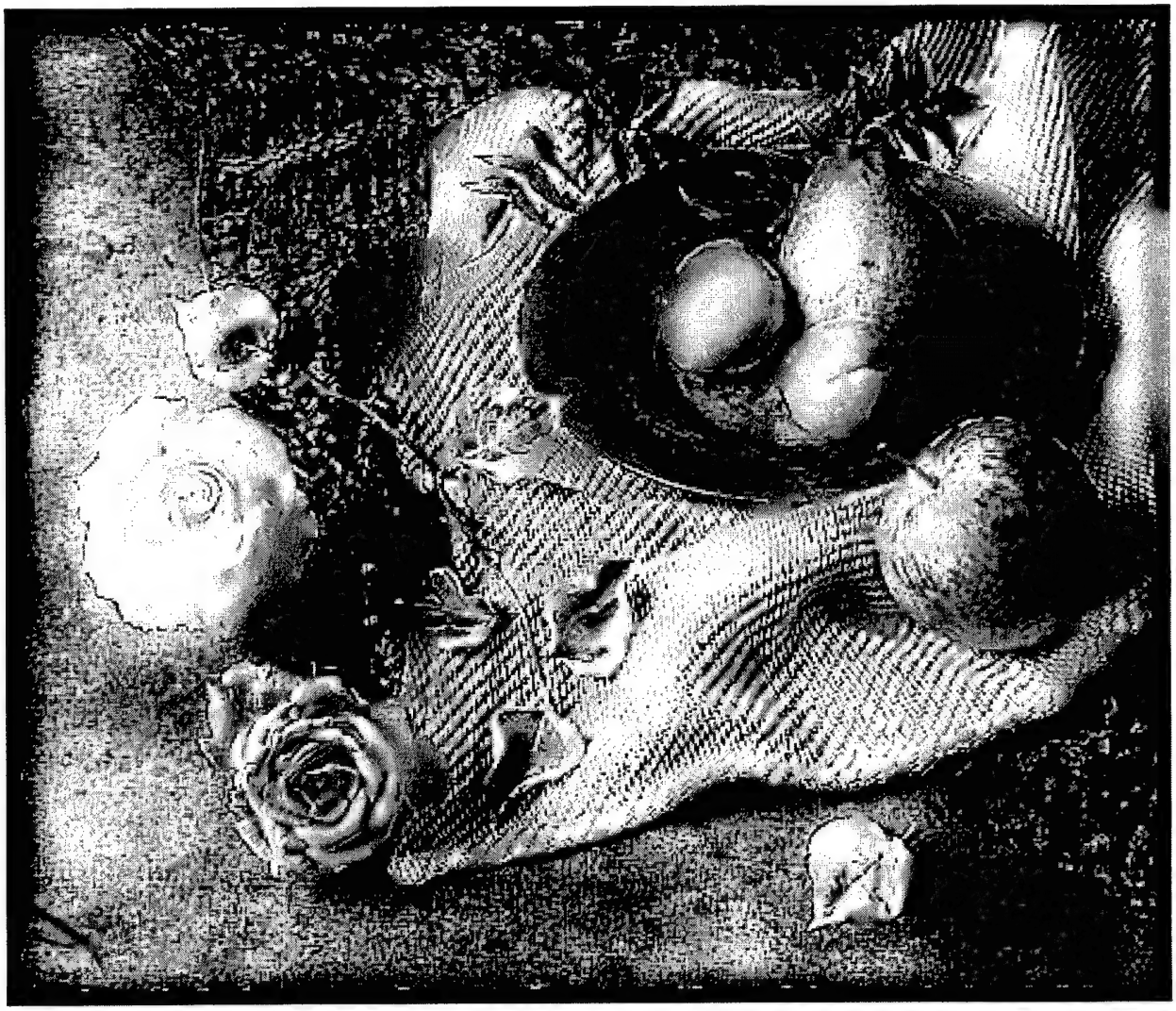

Fig. 29. Example Texture Component Analysis Results

texture channel. Some of the coarse texture appears to have slipped through to the boundary channel. To some extent the contrast boundaries appearing on the texture channel are due to aliasing from having used a uniform kernel rather than a Gaussian kernel in the low-pass filtering operations. Overall however, the algorithm does surprisingly well. The different textures appear clearly on the texture channel and the solid boundaries appear on the boundary channel.

Figure 30 shows the first three planes of the example boundary component analysis pyramid, and figure 31 shows the first three planes of the texture component analysis pyramid. These figures illustrate the pyramids that were rolled-up to create the aggregate images in figures 28 and 29. These images were created by using the zero-one masks produced by the algorithm as "windows" onto the multi-resolution band-pass pyramid of the original image. The equations, which use the multi-resolution texture/boundary segmentation image pyramid, Gf, as a window onto the multi-resolution band-pass pyramid, Bf, and roll-up over all spatial frequencies, are:

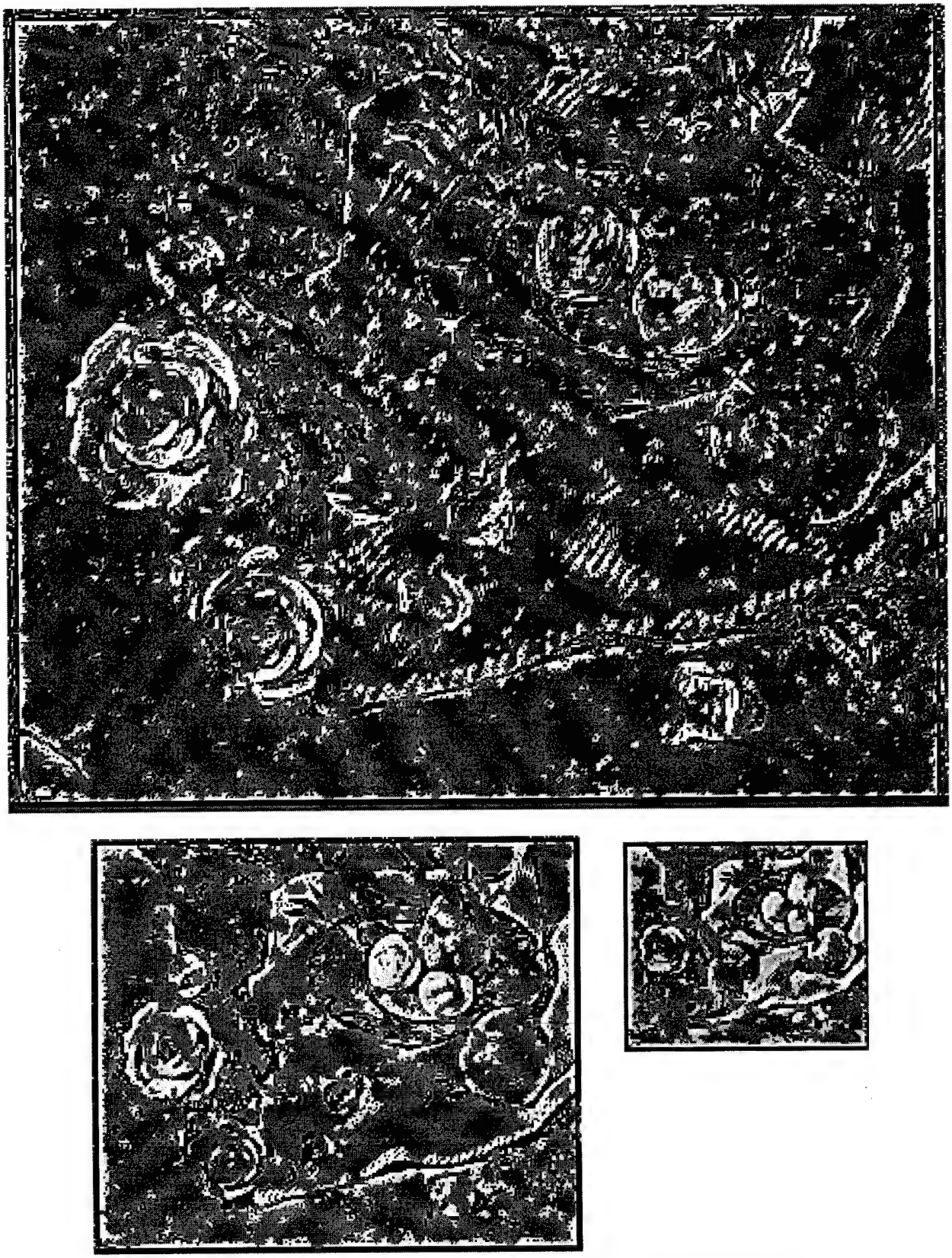

Fig. 30. First three planes of the example boundary component analysis

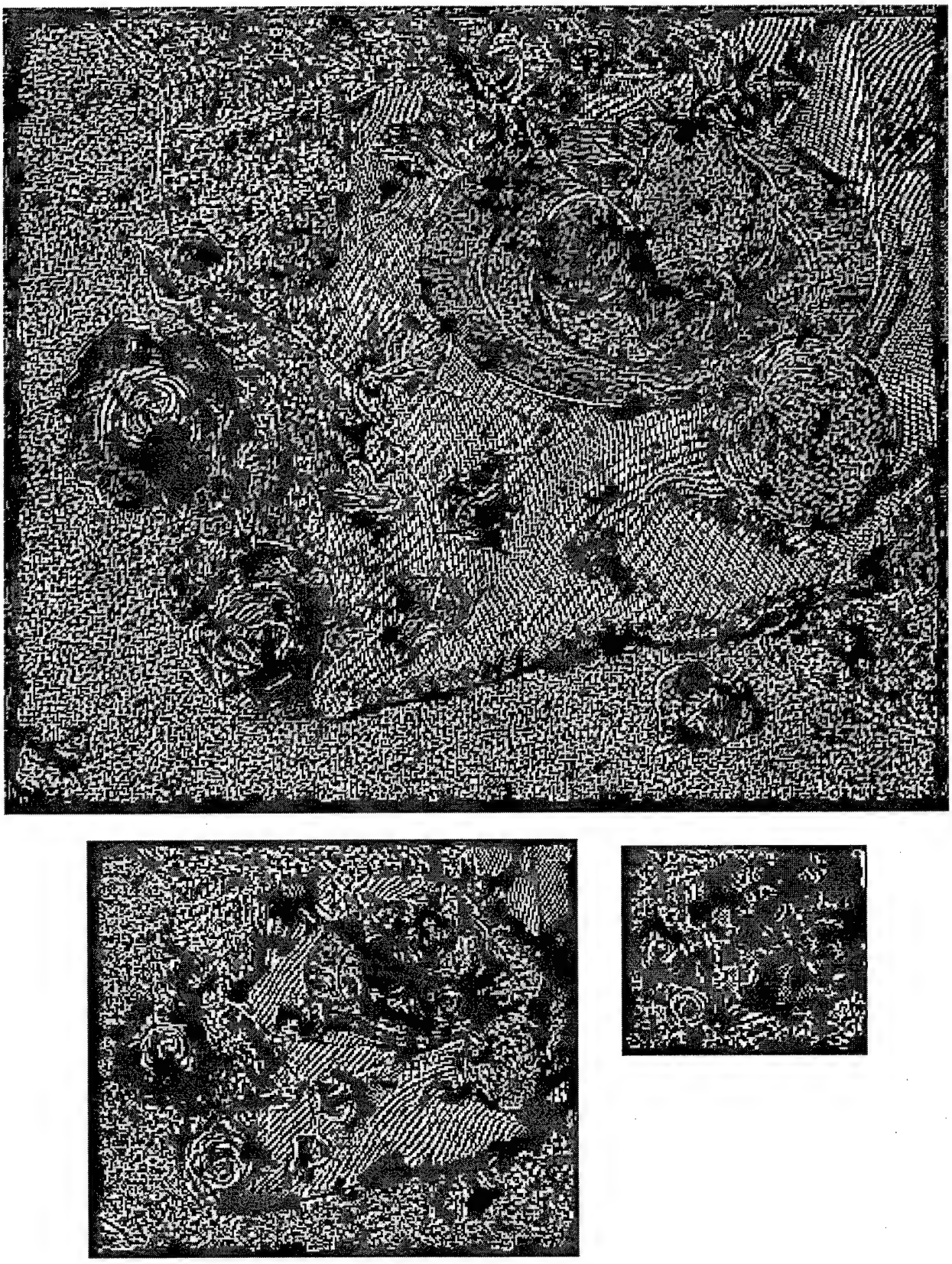

Fig. 31. First three planes of the example texture component analysis

$$W_f = B_f * G_f \quad (33)$$

$$X_{fmax} = W_{fmax} \quad (34)$$

$$X_f = W_f + \text{Expand}(X_{f+1}) \quad (35)$$

$$\text{Boundary Component Analysis Image} = X_0 \quad (36)$$

$$Y_f = B_f * (1 - G_f) \quad (37)$$

$$Z_{fmax} = Y_{fmax} \quad (38)$$

$$Z_f = Y_f + \text{Expand}(Z_{f+1}) \quad (39)$$

$$\text{Texture Component Analysis Image} = Z_0 \quad (40)$$

II.5 Phase II Technical Objectives and Approach

The objectives of this program are to produce calibrated models to predict human visual discrimination response in military scenarios and in commercial automotive scenarios for use in analyzing and evaluating candidate vehicle designs.

The principal Phase II technical objective is to implement, test, evaluate, refine, calibrate, and document models of visual discrimination for three different modes of discrimination and application tasks, per the outline developed in Phase I. The three visual target discrimination models will address: (1) segregating the target from the background and extracting boundary/shape information, (2) matching the target with iconic forms characteristic of alternative target categories, and (3) inferring the nature of complex targets from the spatial and logical relationships among their components. The model development and testing procedure is described in detail in the Work Plan, and the model formulations are described in section 2, above.

The models will be implemented as extensions of the baseline TARDEC VPM, and will be implemented using the TARDEC VPM enabling software workbench. The models will be tested in two stages. The stimuli for development testing will contain alphanumeric characters with various types and levels of degradation. The operational testing to calibrate the models for military and commercial applications will use perception test data from separate Cooperative Research and Development Agreements (CRDAs) between TARDEC and companies in the commercial automotive and military system integration industries. The military testing includes camouflage, concealment, and deception in conjunction with combat vehicles and combat vehicle technologies. The commercial automotive testing includes automotive conspicuity enhancements (e.g., warnings and indicators) and driver's visibility enhancements.

In addition to the principal objective of model implementation and calibration, a supporting objective is to upgrade the underlying VPM enabling software workbench. There are several specific improvements needed to transform it from "research-grade" software to "commercial-grade" software. The specific details of these improvements are described in the Phase II Work Plan. To the maximum extent possible, the software upgrades will make use of commercial subroutine libraries.

This section outlines the approach to model implementation, test and evaluation in Phase II. The major events are the annual software demonstrations and deliveries, with the accompanying documentation in the annual technical report. The final product will consist of three models of visual discrimination, tested and calibrated against empirical perception test data for military and commercial automotive applications. The models will be demonstrated in the context of the military and commercial automotive applications to illustrate the transition of the products to the Government and private sectors. The technical work is organized into two parallel tasks: (1) target discrimination model implementation and calibration, and (2) VPM enabling software workbench upgrades.

II.5.1 Target Discrimination Model Implementation and Calibration

The model implementation, test, and calibration is organized into three sequential stages. Each stage address a different aspect of visual discrimination modeling, as determined in the Phase I research.

The first subtask is to implement, test, and calibrate a predictive model of target segregation and shape discrimination based on an information metric derived from the visibility of the target boundary and a measure of the complexity of the target shape. The objective of this model is to predict how well observers are able to distinguish the target from the background and determine its shape. This model will not attempt to predict what target classification decision people make, only their accuracy in making the correct decision.

The second subtask is to implement, test, and calibrate a model to predict what target classification decision observers will make, using a pattern-matching methodology and a set of icons representing the characteristic shapes of the different target types. Both the first and second subtasks are restricted in scope to simple target characterizations in which the external shape of the target is sufficient for discrimination, but in which logical deduction based on discrimination of internal components is not required.

The third subtask builds on the results of the preceding work to model the discrimination of complex targets based on the spatial and logical relationships among individually discriminated component parts.

Each of these subtasks will be accomplished in three steps: (1) implementation, (2) character-recognition testing and model refinement, and (3) calibration to application-task data.

II.5.1.1 Target Discrimination Model Implementation

The first step is to implement the initial formulation and any significant alternatives on the VPM enabling software workbench. The implementations will build upon the existing modules and will interface with the front-end VPM. The details of the initial model formulations are described in section 2, Phase I Results.

II.5.1.2 Character-Recognition Testing and Model Refinement

The second step is to test and refine the model using perception test results for alpha-numeric character recognition, in the presence of noise, low contrast, clutter, obstruction, and other degraded conditions and transformations. We will compare the model predictions to the results of human-observer perception testing. The rationale for using character recognition for the internal screening testing includes the following:

- Alpha-numeric character recognition is a standard and well-accepted approach in vision research, and is widely used to measure visual function (e.g., standard Snellen charts for visual acuity, Ishahara color-blindness charts, Pelli-Robinson low contrast letter charts, etc.).
- The model calibration for alpha-numeric character recognition testing will be applicable to the design and evaluation of console displays in military and commercial vehicles.
- Character-recognition testing is simpler than testing with complex vehicle scenarios, and thus can more easily be performed with rigorous scientific control and experimental design, and can be accomplished with an economy of time and resources.

- Character-recognition testing is less subject to variations due to subjective interpretation and learning effects than are automotive and military discrimination tasks and scenarios.
- The test patterns and perception test results may be valuable in and of themselves to diagnose or evaluate aspects of visual performance not addressed with standard eye charts.

A central element of model development will be iterative test and refinement with character-recognition test stimuli. This will be done prior to calibration to the application-based test data. By applying different levels, types, and mixes of degradation, we can ensure that the models are robust, or determine the region in which they are accurate, and the region in which they are not. The use of character charts enables us to test over a wide variety of signature scenarios, with a scientifically controlled experimental design and well-established test protocols. This scientific rigor would not be possible with the application-based test data.

The test stimuli will be pre-tested to ensure a range of subject responses, i.e., to ensure that some cases are hard, some are easy, and some are in between. The stimuli must be designed for significant variance in the subject response over the control variables, so that we can statistically test the ability of the model to explain the variance in subject response.

Figures 32 to 36 illustrate some of the types of methods for generating the character-recognition charts which may be used. These charts are presented as illustrations only. Their purpose is to illustrate some of the variety of image scenarios for developmental testing. The specific stimuli will be generated to provide data for the tasks and conditions most appropriate at each stage of modeling.

We have yet to determine whether we will use the standard eye chart approach in which character size is varied in every chart and other control parameters are varied between charts. An alternative approach is to hold all factors (including size) constant in each chart. Standard eye charts support collecting accuracy data but not response time data. The alternative approach is better suited to collecting both accuracy and response time data. Accuracy data can include percent correct identification (what the character is), and percent correct recognition (that there is a character there), and percent correct detection (that something other than background is present).

The characters in figure 32 were created by changing the following image parameters: size, noise magnitude, target contrast, target edge sharpness/blur, target boundary ripple/spatial modulation, edge contrast, and texture contrast. Figure 33 illustrates degraded signatures created by obstructing the targets at different density and spatial frequency. Figure 34 shows rotation deformations, and figure 35 shows targets created by texture (spatial frequency and orientation) gradients. Figure 36 shows the same character distorted in different fonts (out of context, some of these characters would be difficult for a layman to recognize as an English letter or to identify as the letter "A"). These figures are presented for illustration purposes only. The test stimuli may combine one or more of theses methods, as appropriate for the stage of model testing.

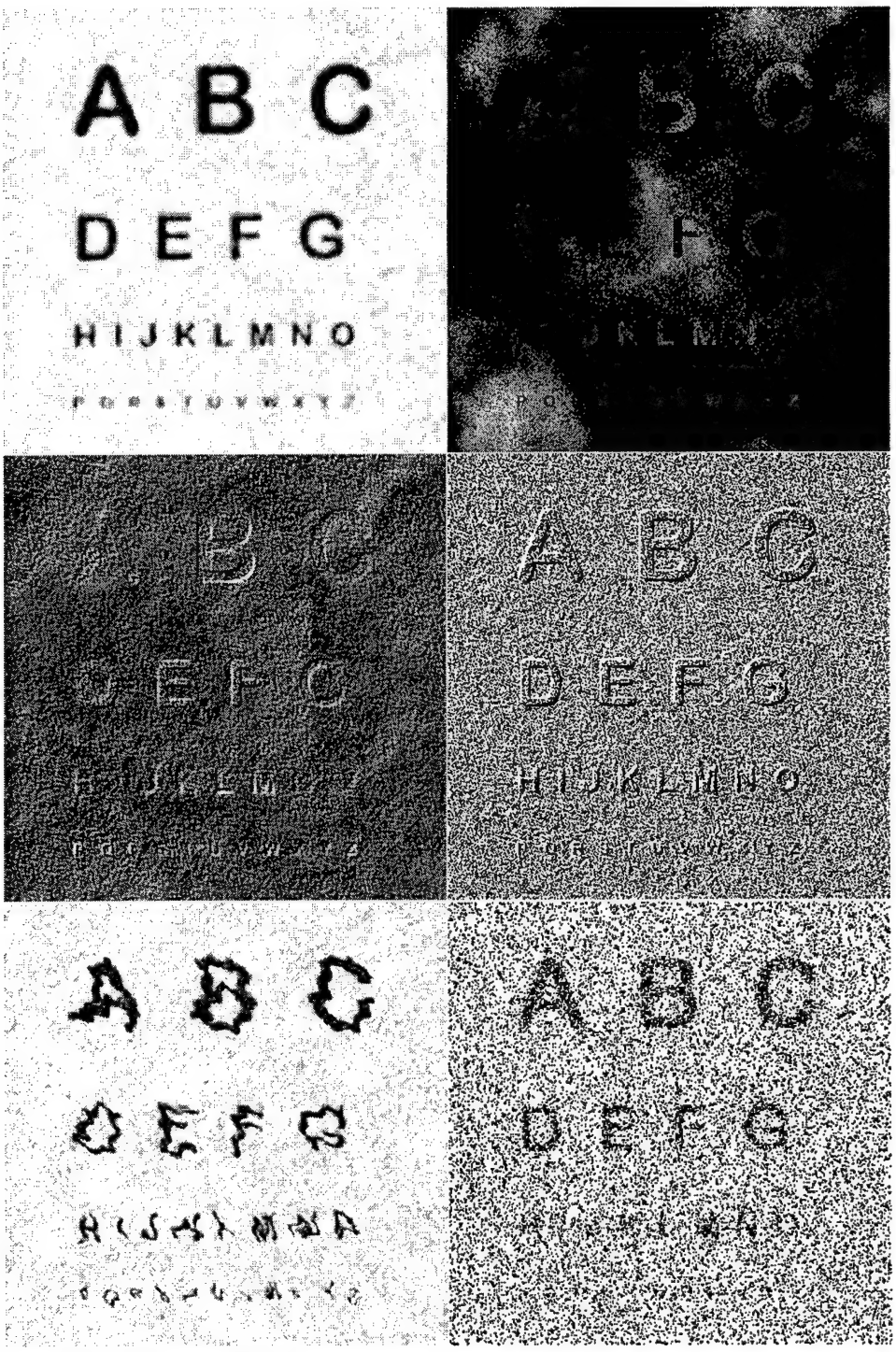

Fig. 32. Example image degradation methods for character-recognition stimuli

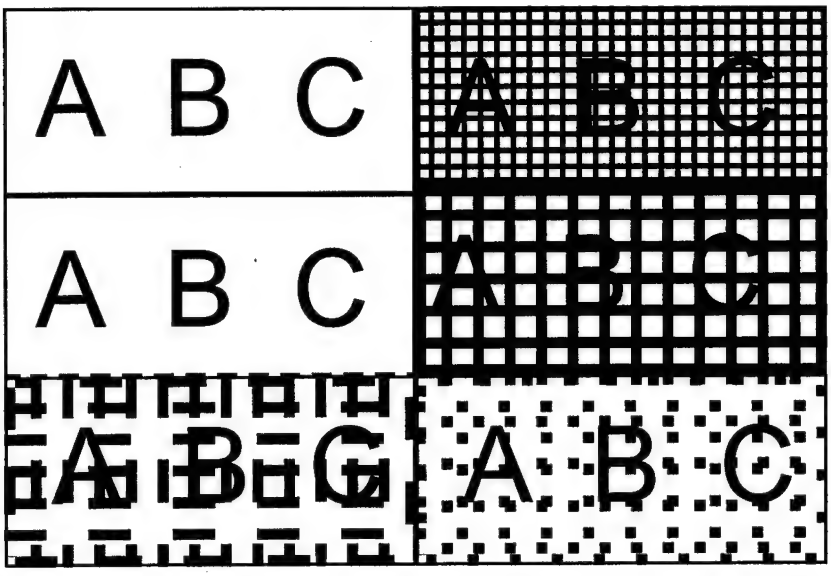

Fig. 33. Example obscuration methods for character-recognition stimuli

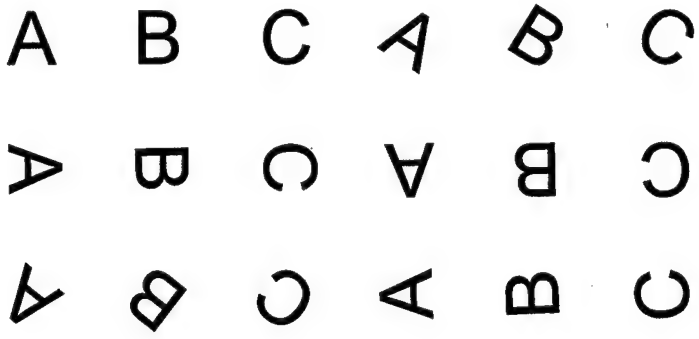

Fig. 34. Example rotation deformations for character-recognition stimuli

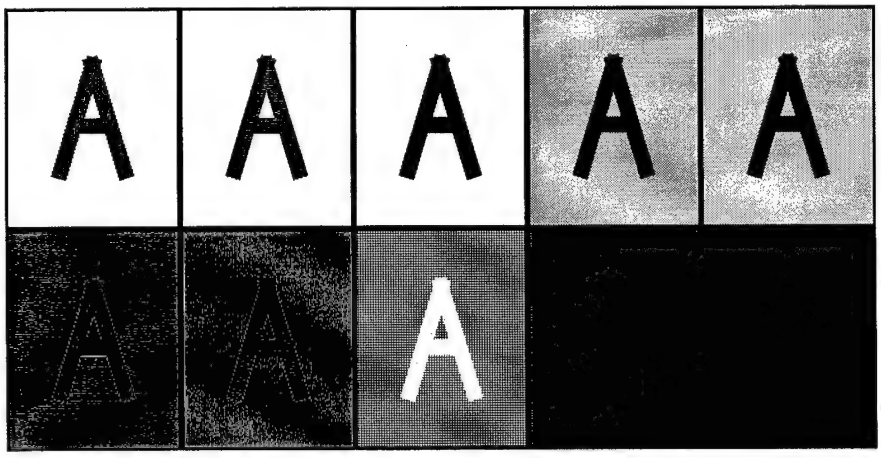

Fig. 35. Example texture-contrast methods for character-recognition stimuli

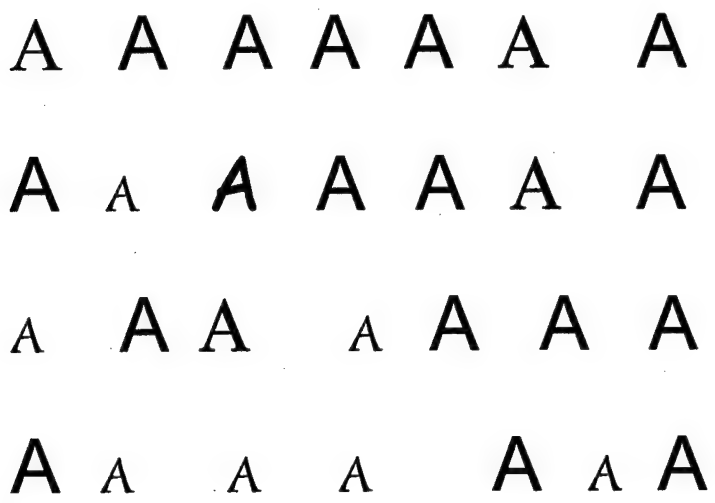

Fig. 36. Examples of different fonts for the same character

II.5.1.3 Calibration to Application Task Data

The third step is to test and calibrate the model using perception test data collected in complex vehicle scenarios. Data for this testing will be generated as part of the CRDAs between TARDEC and the military system integrators, and between TARDEC and the automotive manufacturers. The complex vehicle perception testing is not part of the Phase II SBIR project. Turing Associates, Inc., will support TARDEC in the CRDAs and participate in the design, execution, and analysis of the tests. However, this participation in the CRDAs is not part of the SBIR project; this content is the related activity outside the scope of the SBIR contract which is being paid for by the TARDEC matching funds. The results of the step-three model test and calibration will be coordinated with the CRDA partners to ensure acceptance and applicability. This aspect of the model development and testing is focused on ensuring a high probability of successful commercialization by making sure that the modeling is focused on Government and industry design priorities, and that the models are calibrated and demonstrated in the context of high-priority design applications.

II.5.2 VPM Enabling Software Workbench Upgrades

The visual perception models are implemented as data flow diagrams on top of VPM enabling software workbench execution engine. The data flow diagrams connect native modules of the workbench written in C++, and lower-level data flow diagrams. This provides a modular hierarchical structure for implementing complex modules with an economy of effort, maximum re-usability of component modules, and minimal low-level (C++) coding. The data-flow-diagram approach provides a highly visual presentation of the model, and the hierarchical structure simplifies the picture at each level.

The VPM enabling software workbench consists of two components. The first component reads in the data flow diagrams, instantiates the hierarchical modules, and flattens the entire network to a complex data flow diagram among only the native modules. The second component instantiates the native modules, links them, and executes the model. The workbench automatically handles memory management and inter-process communication. The workbench is referred to as the "enabling" software because without it the visual perception models could not be executed.

There are several enhancements to the VPM enabling software workbench which will greatly improve its efficiency, and which are needed to transform it from "research-grade" software to "commercial-grade" software. The workbench upgrade is organized into two tasks. The first task is to upgrade the infrastructure. The second task is to implement a graphic user interface (GUI) to display, edit, and create data flow modules and applications.

II.5.2.1 Infrastructure Upgrades

There are three significant infrastructure upgrades: (1) memory management improvement, (2) configuration management improvement, and (3) image-display and region-editing interface improvements.

II.5.2.1.1 Memory Management Improvement

We propose to implement automatic caching and release at output ports without explicit caching control statements in the data flows. The current version of the TARDEC VPM enabling software workbench requires that explicit cache and release controls be programmed by the model developer. The improved memory management will reduce execution time, and improve the clarity of the data flow specifications.

II.5.2.1.2 Configuration Management Improvement

Configuration management improvement will be accomplished by separating development and application versions of the VPM enabling software workbench. Currently, the workbench recursively reads a set of hierarchical data flow diagrams, instantiates the modules, and creates links between the modules until all modules are resolved down to the lowest level of executable native data types. Then the workbench executes the model. We propose to split this process into two steps: (1) create the executable model from the data flow diagrams, and (2) execute the model. We propose to implement two versions of the VPM enabling software workbench: (1) a model development version which has the full functionality of the VPM enabling software workbench and adds the option to output a flat file of the data flow with all hierarchical modules resolved down to connections between native data types (i.e., to collapse the hierarchical form of the model into the flat, fully-resolved form of the model), and (2) an execute-only version for turn-key applications which reads and executes the fully-resolved form of the model, but which cannot resolve the hierarchical form of the model. These will be called the Developer's Workbench and the Application Execution Workbench. At the present time, the only version is the development version. This hinders configuration management because any user could, either accidentally or intentionally, modify the model. Having an "execute-only" version will be faster for turn-key applications, and will eliminate potential configuration management problems.

II.5.2.1.3 Image-Display and Region-Editing Interface Improvements

Improved image editing and display will utilize commercial software libraries. At the present time, the VPM enabling software workbench has limited capability to display images and intermediate model outputs or to edit images to designate target regions. We propose to integrate display and editing routines for standard operations, and to improve the format conversions to export and import image data.

II.5.2.2 Graphic User Interface

The GUI allows the user to create, edit, and display data flow applications and hierarchical modules. Currently the data flow applications and models are specified as text files. The user has to manually translate the data flow diagrams into text files, typing the name of the modules and input-output port connections. More importantly, to document the model, the user has to manually draw data flow diagrams from the connections files. These processes are time consuming and error prone. The solution is to implement a GUI so that the user can simply select modules from the list of modules in the library, "drag and drop" icons for the modules, draw the connections between the input and output ports, add hierarchical data flow modules to the library, display embedded documentation comments, and "zoom in" on hierarchical modules to display the lower-level map. This eliminates the need for the user to remember or look up the module and port names. The GUI will automatically check the data type match between the output port of one module and the input port of the next. This task will use commercial software libraries to build GUIs.

APPENDIX A: REFERENCES

Bergen, J. R. and M. S. Landy

1991 Computational Models of Visual Texture Segregation. Pp. 253-71 in M. S. Landy and A. J. Movshon, eds., *Computational Models of Visual Processing*. Cambridge: MIT Press.

Biederman, I.

1987 Recognition-by-Components: A Theory of Human Image Understanding. *Psychological Review* 94:115-47.

Brady, M. and A.Yuille

1987 An Extremum Principle for Shape from Contour. Pp. 329-60 in M. A. Arbib and A. R. Hanson, eds., *Vision, Brain and Cooperative Computation*. Cambridge: MIT Press.

DeValois, R. L. and K. K. DeValois

1990 *Spatial Vision*. Oxford: Oxford University Press.

Feldman, J. A.

1987 A Functional Model of Vision and Space. Pp. 531-61 in M. A. Arbib and A. R. Hanson, eds., *Vision, Brain and Cooperative Computation*. Cambridge: MIT Press.

Geisler, W. S.

1989 Sequential Ideal-Observer Analysis of Visual Discriminations. *Psychological Review* 96.2: 267-314.

Gonzalez, R.C., and Wintz, P.

1987 *Digital Image Processing*. Reading MA: Addison-Wesley Publishing.

Graham, N. V. S.

1989 *Visual Pattern Analyzers. Oxford:* Oxford University Press.

Kosslyn, S. M.

1994 *Image and Brain*. Cambridge: MIT Press.

Kosslyn, S. M., C. F. Chabris, C. J. Marsolek, and O. Koenig

1992 Categorical Versus Coordinate Spatial Representations: Computational Analyses and Computer Simulations. *Journal of Experimental Psychology: Human Perception and Performance* 18:562-77.

Lowe, D. G.

1987a Three-Dimensional Object Recognition from Single Two-Dimensional Images. *Artificial Intelligence* 31: 355-95.

1987b The Viewpoint Consistency Constraint. *International Journal of Computer Vision* 1:57-72.

1985 *Perceptual Organization and Visual Recognition*. Boston: Kluwer.

Marr, D.

1982 *A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: Freeman Press.

Medin, D. L. and Ross, B. H.
1991 *Cognitive Psychology*. New York: Harcourt Brace Jovanovich, 1991.

Nakayama, K.
1990 The Iconic Bottleneck and the Tenuous Link between Early Visual Processing and Perception. Pp. 411-22 in C. Blakemore, ed., *Vision: Coding and Efficiency.* Cambridge: Cambridge University Press.

Rueckl, J. G., K. R. Cave and S. M. Kosslyn
1989 Why Are "What" and "Where" Processed by Separate Cortical Visual Systems? A Computational Investigation. *Journal of Cognitive Neuroscience* 1:171-86.

Schmucker, K. J.
1984 *Fuzzy Sets, Natural Language, Computations, and Risk Analysis*. Rockville MD: Computer Science Press.

Singh, H., Meitzler, T., Gerhart, G., Arafeh, L.
1996 Fuzzy Logic Approach for Computing the Probability of Target Detection in Cluttered Scenes. *Optical Engineering* 35.12:3623-36.

Standing, L.
1973 Learning 10,000 Pictures. *Quarterly Journal of Experimental Psychology* 25:207-22.

Wandell, B.
1995 *Foundations of Vision*. Sunderland MA: Sineauer Associates.

Witus, G.
1996 *TARDEC National Automotive Center Visual Perception Model, Final Report: Analyst's Manual and User's Manual*. OMI-577, prepared under contract DAAE07-94-C-R111. Ann Arbor: OptiMetrics, Inc.

Wolfe, J. M. and S. C. Bennett
1997 Preattentive Object Files: Shapeless Bundles of Basic Features. *Vision Research* 37.1:25-43.

Zipser, D. and R. A. Anderson
1988 A Back-Propagation Programmed Network that Simulates Response Properties of a Subset of Posterior Parietal Neurons. *Nature* 331:679-84.

APPENDIX B: VPM EXTENSION MODULES FOR ALGORITHM DEMONSTRATIONS

The following modules were created as part of this project to demonstrate selected algorithms. The source code for the modules has been provided to TARDEC on magnetic media. These modules run on the TARDEC VPM workbench.

aRun2ndStage.txt
aRunDeadband.txt
aRunGains.txt
aRunMRMask.txt
aRunNoTarget.txt
aRunOneMinusCDP.txt
aRunSegBndry2.txt
aRunSegmentor2.txt
aRunShapeMetric.txt
aRunTextureSeg.txt

APPENDIX C: KOSSLYN'S ANALYSIS OF VISUAL COGNITION

Kosslyn's analysis focuses on "high-level" cognitive processing in visual recognition and identification, i.e., processing in the context of previously developed characterizations and expectations of the appearance and relationships of objects and events. He conceives of visual cognition as being accomplished by a network functionally independent subsystems, i.e., other than providing inputs, the functioning of one subsystem does not affect the functioning of any other subsystem. The subsystems are "plug-compatible". The subsystems operate in parallel, not series, with as much feed-back as feed-forward resulting in true cooperative computation. The subsystems tend to have relatively specific anatomical locations, but may have overlapping implementation (i.e., one anatomical location may participate in more than one subsystem).

Kosslyn's model of the brain's visual cognition processing architecture is illustrated in figure 37. Figure 37 is a reproduction of figure 11.1 in Kosslyn [1994: 383]. The input, process, output, and anatomical location in the brain of each function are summarized in table 11.1 in Kosslyn [1994: 380-2]. Kosslyn's book is devoted to developing the model, and presenting evidence for each function. Kosslyn seeks convergent evidence from psychophysical tests of normal and brain damaged humans, and combined pyschophysical/neurological tests of normal humans, brain damaged humans, and animals.

Kosslyn's model is a functional model, not a computational model. No algorithms are presented. Kosslyn uses computational analyses to illustrate the types of algorithms which could be at work in visual cognition. The primary image processing algorithm methods are multi-resolution band-pass analysis, correlation for visual pattern matching (including whole-figure matching and matching to canonical part), and Baysean classification and neural nets for object categorization. He draws on the work of Biederman [1987], Lowe [1985, 1987a, 1987b], Marr [1982], Zipser and Anderson [1988], his own original research [Kosslyn et al. 1992], and others.

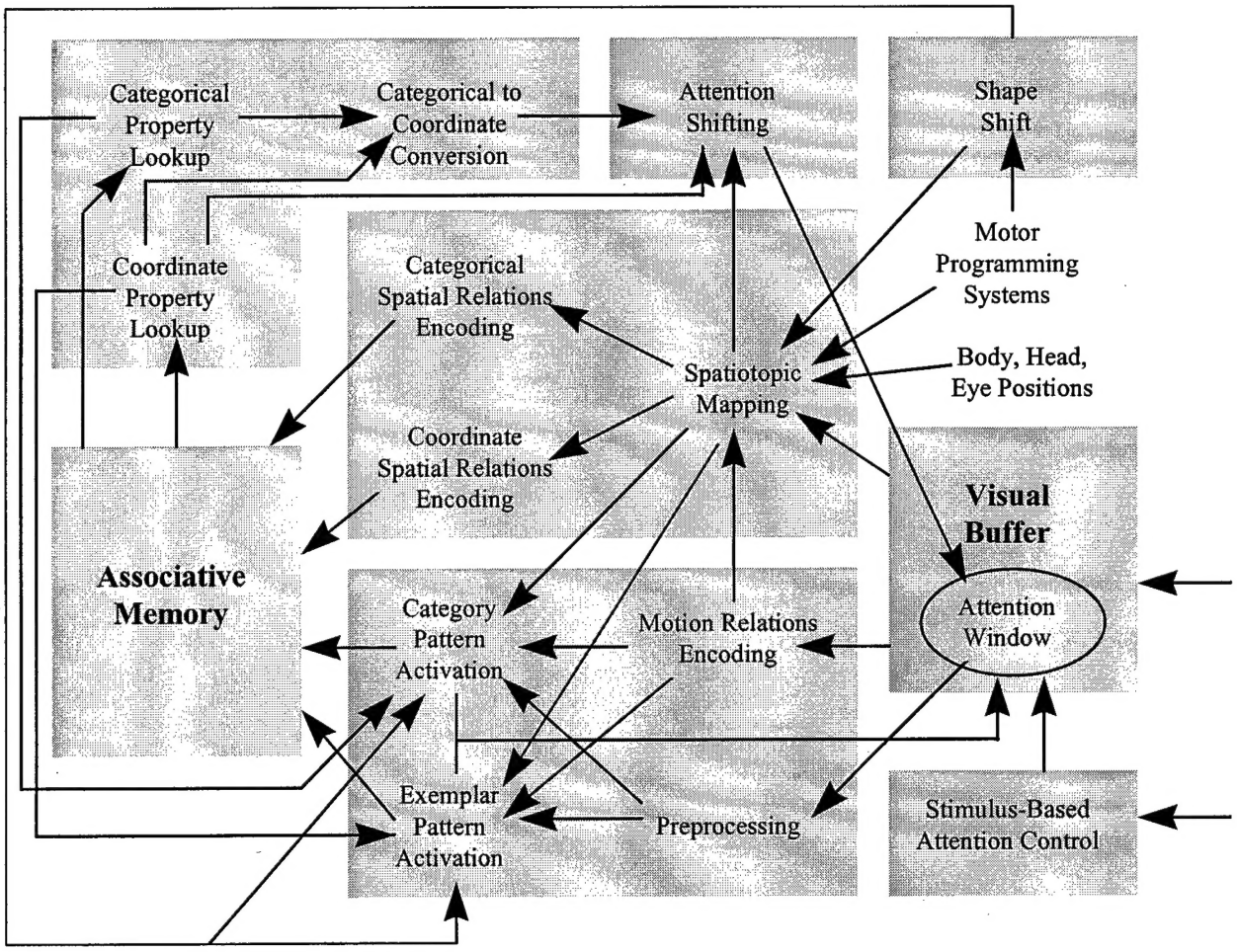

Fig. 37. Kosslyn's architecture of image processing